Notes of this revised deck:
The following deck is a revised presentation deck, for use in the video interview on 26th April 2021.

This deck now includes the following:

- New slides of analysis to address topics not convered in the initial submission, including: technical details of Logistic Regression and XGBoost, their strengths and weaknesses, metric selection and model calibration, as well a detailed slides of each key shot. These slides feature an orange diamond in the top-right corner - ♦
- A second Chance Quality Model created using XGBoost (see slide 25 and Appendix), building on the Logistic Regression model in the initial submission. More information about this model can be found in following notebook [link].

Junior Data Science Challenge

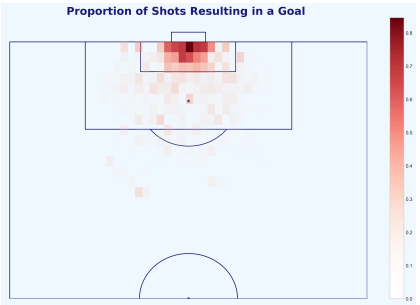
Edward Webster

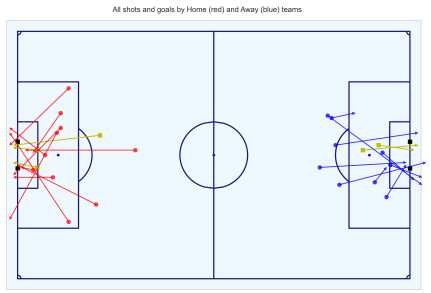
April 2021

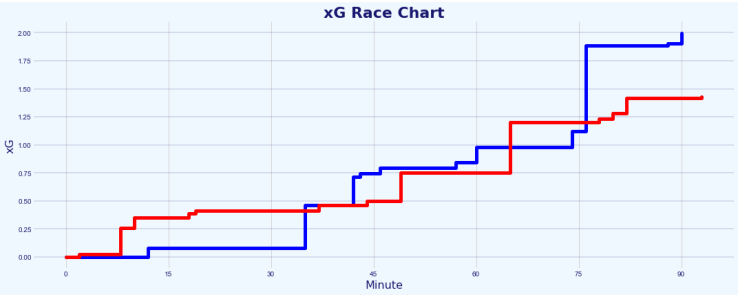
Slides: https://docs.google.com/presentation/d/116D0U_ue2sv6hgLBgHnil228cRhph0qfMuOFA4uuxhs/edit?usp=sharing
Code: https://github.com/eddwebster/mcfc_submission
Full pack including all attachments: https://drive.google.com/drive/folders/1Ts_5YOL8JpVpRcaEhqmcTCceLmcBRY3L?usp=sharing

City football METRICA SPORTS

Contents







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

Introduction and Challenge Brief

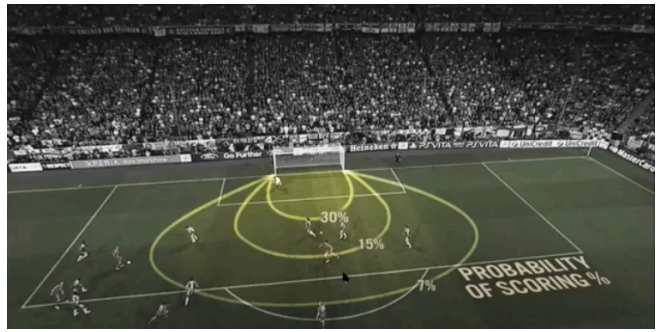
Brief

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

1

Building a Chance Quality Model using shots data

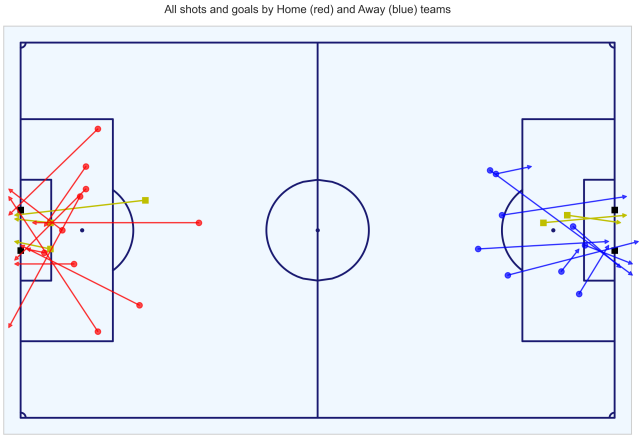
Use the provided shot data that build a Chance Quality Model (CQM) that calculates the probability of a shot resulting in a goal.



2

Analysis of the Metrica Sports Sample Data

Analyse the shooting opportunities for each team in game two of the sample Metrica Sports Tracking and Event data, making use of the trained CQM, to determine which team deserved to win the game?



Expected Goals images taken from both David Sumpter's Friend of Tracking seminar: 'How to Build An Expected Goals Model 1: Data and Model (https://www.youtube.com/watch?v=bpjLyFyLIXs).
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.

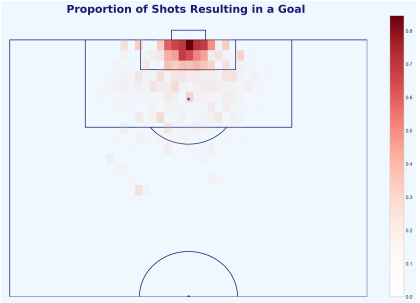
Analysis Process

Explanation of how this data challenge was tackled in three parts

1

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

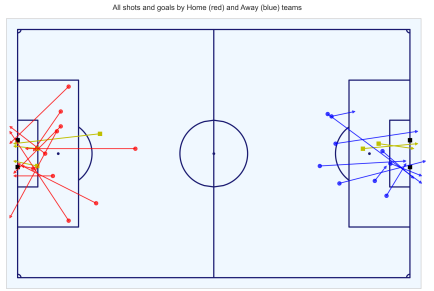
Build and train a basic classification model that can predict the likelihood of shots resulting in goals



2

Analyse game 2 of the sample Metrica Sports Tracking and Event Data

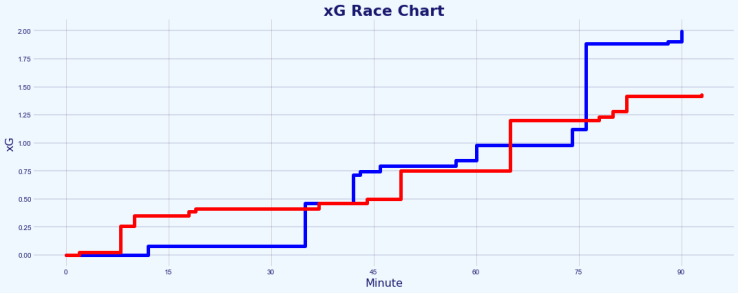
Using the Tracking data and corresponding Event data, examine the key chances in the match and export a shots dataset compatible to be used with the trained Chance Quality Model.



3

Assessment of the team's performance in the Metrica Sports data through application of the Chance Quality Model

Using the exported Metrica shots data to predict the likelihood that the chances that took place in the match resulted in a goal or not through application of the trained Chance Quality Model.



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

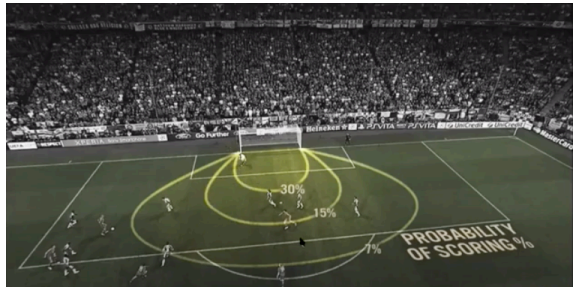
Chance Quality Model

Key objectives achieved with the Shots dataset, summarised in this section

1

What are Expected Goals (xG)?

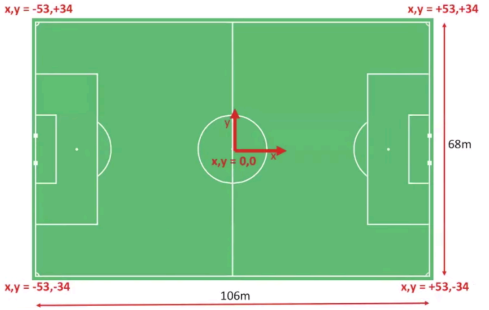
Background information about the proxy of choice for the Chance Quality Model. What is this xG metric and why has it made such a huge impact on football data analysis.



2

Data Engineering

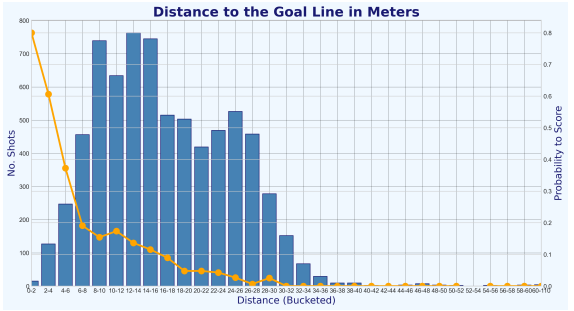
Reworking the dataset into a form ready for modeling including: converting the pitch coordinates to a standardised coordinate system, filtering the data for just Open Play goals, and cleaning attributes.



3

Building a Chance Quality Model *via* the logic of Expected Goals

A basic model to measure the likelihood of shots resulting in goals, steps include: exploratory data analysis (EDA) of the dataset, data engineering, model selection evaluation metric selection, outlier treatment, univariate and multivariate analysis with subsequent feature engineering, production of a final model, measurement of performance, and feature interpretation.



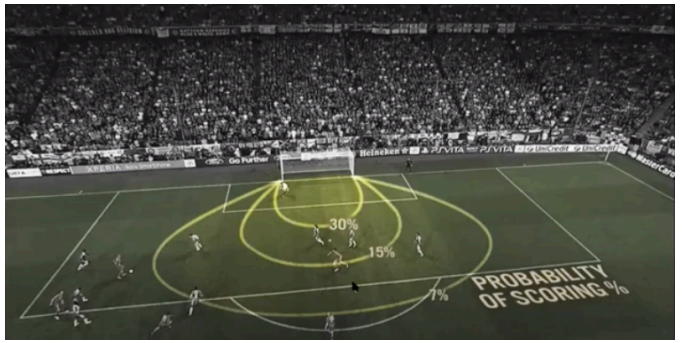
Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

What are Expected Goals and Why is this an Important Metric?

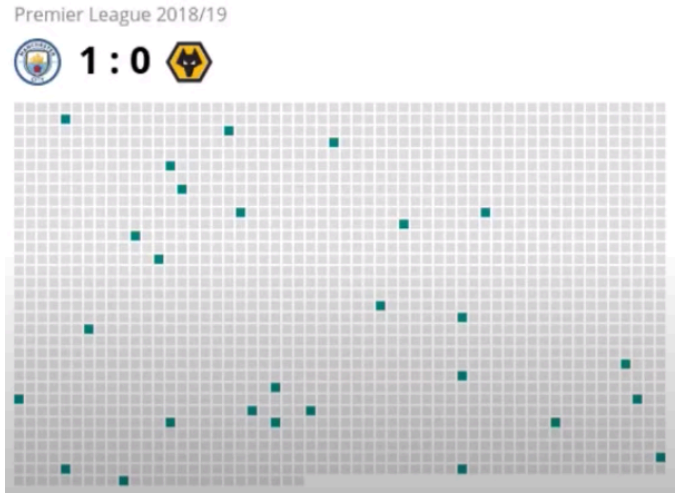
Arguably the most important metric in football analytics so far – Expected Goals

- 'Expected goals', often abbreviated as 'xG', are the probability that a shot with defined state will, on average, result in a goal (Fig. A). This principle of xG is what is used to create the Chance Quality model.
- Shots in matches are rare and goals are even rarer! On average, each team shoots 12 times per matches (0.4% of a 3,000 Event dataset) resulting in 2.66 goals (Fig. B).
- xG 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.
- xG can be used to make smarter decisions on recruitment, tactics, and strategy. For example, when recruiting a striker in good form, xG can be used to see if the player is over- performing or is in fact taking the chances that would be expected.

A



B



Definition of Expected Goals derived from Laurie Shaw's article: Bodies on the Line: Quantifying how defenders affect chances: http://eightyfivepoints.blogspot.com/2017/09/bodies-on-line-quantifying-how.html.
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.
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

Exploratory Data Analysis of the Raw Dataset

Information about the provided Shots dataset of just under 11,000 shots used to train the Chance Quality Model

- The shots DataFrame contains 10,925 shots (Fig. A) and 1,374 goals (12.6%) (Fig. B)

- The key features available:
  - X and Y coordinates of the shot.
  - The Play Type i.e. the game situation in which the shot was taken (open play, penalty, direct free kick, direct from a corner);
  - The Body Part which shot was taken (left foot, right foot, head, other);
  - The No. of Intervening Opponents and Teammates i.e. the number of players that were obscuring the goal at the instant of the shot (from the perspective of the shot-taker);
  - The Interference of the Shooter i.e. the pressure the shot-taker is experiencing 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); and
  - The shot Outcome i.e. the result of the shot (blocked, missed, goal frame (post or bar), saved, goal or own goal).

< 11,000 Shots in the Dataset

A

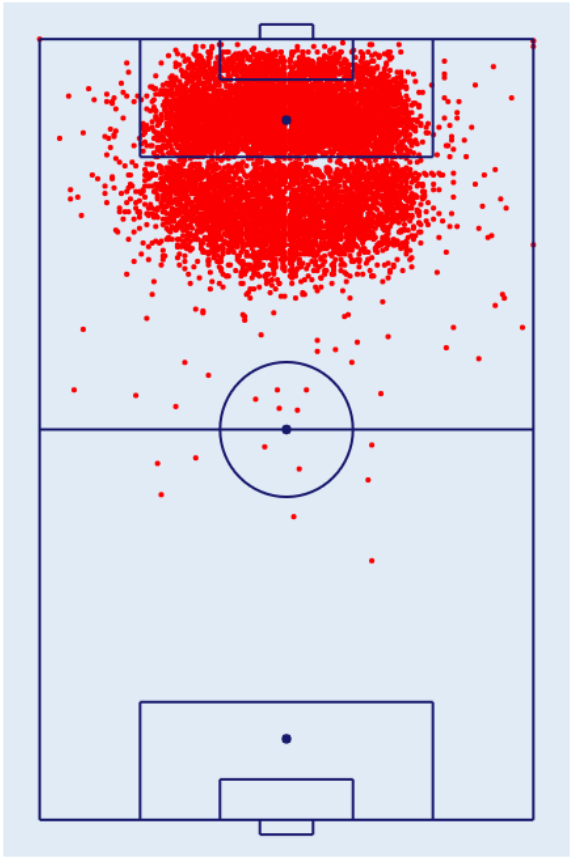


Fig. A: All shots in the dataset.

B

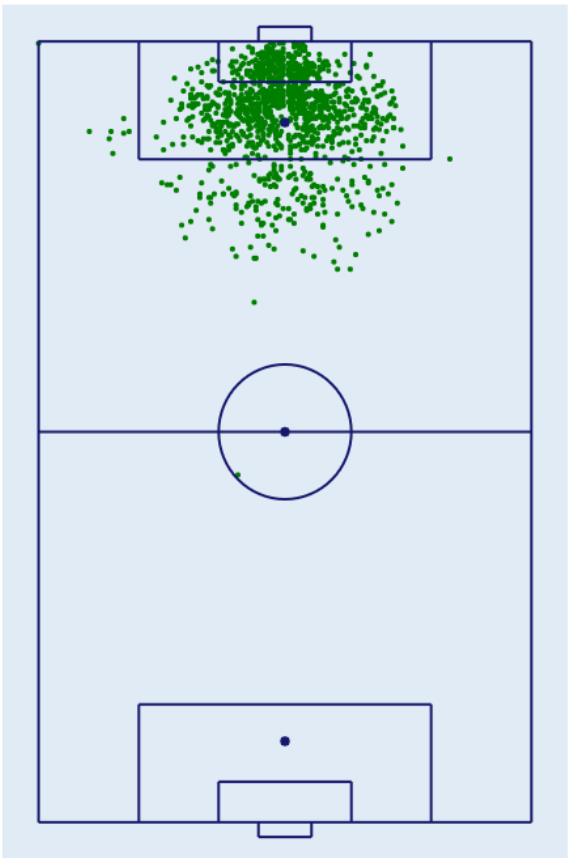


Fig. B: All goals in the dataset.

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: https://github.com/eddwebster/mcfc_submission/blob/main/documentation/shots/ShotData.txt

Initial Data Engineering

Steps to engineer the dataset before modeling – coordinate conversion, data cleaning, and subsetting the data

- **Conversion of Pitch Coordinates** – to match those used by Laurie when working with the Metrica Sports data (see GitHub repository [link]).
  - Original: x(+106, 0)m, y(-33.92, +33.92)m (Fig. A)
  - Converted: x(-53, +53)m, y(-34, +34)m (Fig. B)

- **Data cleaning** – replace NULL values for 'Interference_on_Shooter with 'Unknown' string and values in the 'play_type' column tidied to be just 'Direct Corner'.

- **Removal of own-goals** (43)

- **Define target variable, 'isGoal'** - whether the shot was a goal (1) or not (0), derived from the 'Outcome'.

- **Subset for only Open Play shots** - removing all shots from penalties, free kicks, or directly from corners

- **Overall** - in total, 699 shots (6.4%) and 199 goals (14.5%) were removed from the dataset.

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

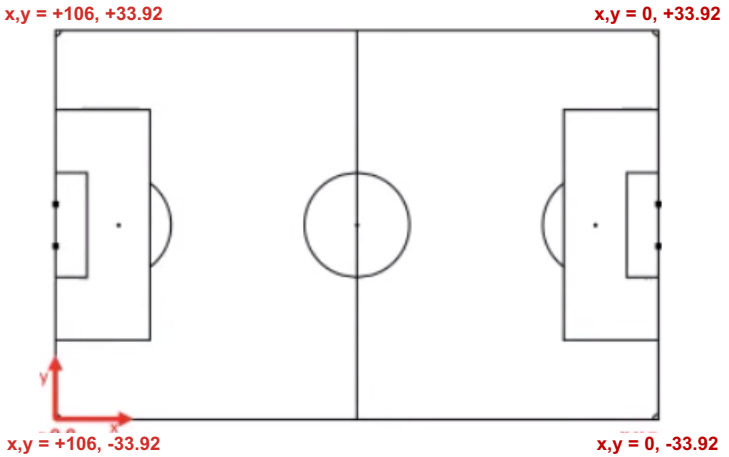


Fig. A: Pre-conversion of coordinates

B

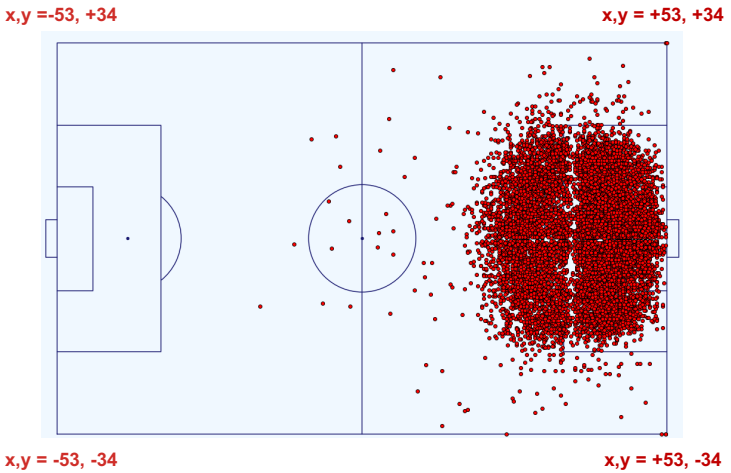


Fig. B: Post-conversion of coordinates including shots

Initial Modelling

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

- **Classification model selection** – Logistic Regression (and subsequently XGBoost). See the Appendix and or the individual model notebooks (Logistic Regression [link] and Gradient Boosted Decision Trees [link]) for more information.

- **Initial model considerations:**
  - Baseline for comparison to future iterations
  - Training and Test split – 0.7:0.3
  - Feature variables – only used standardised X and Y coordinates and Body Part
  - Target variable – isGoal
  - Future iterations of the model include engineered features and one-hot encoded categorical features.

- **Evaluation metric selection:** Log Loss (Binary Cross Entropy). See the Appendix and CQM notebook [link] as to why Log Loss is the most situation metric for this binary classification problem.

Evaluation Metric	Value
Accuracy	88.5%
Log Loss*	0.33182
ROC AUC	72.8%

Primarily Log Loss and then also ROC AUC were the evaluation metrics of choice for evaluating the performance of the model.
Logistic Regression wiki: https://en.wikipedia.org/wiki/Logistic_regression.
XGBoost official documentation: https://xgboost.readthedocs.io/en/latest/.
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

- A number of Open-Play shots taken and subsequently scored from improbable locations on the pitch (Fig. A). Most likely due to uncharacteristic errors from the goalkeeper (see Xavi Alonso example below and [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. These goals therefore excluded by changing the result of all shots with distance >35m and all shots from >18m with shot angle >35 degrees removed from the data.
- This results in 33 Open Play shots values having their value changed from goal (1) to no goal (0) - 0.27% of the Open Play shots.

A

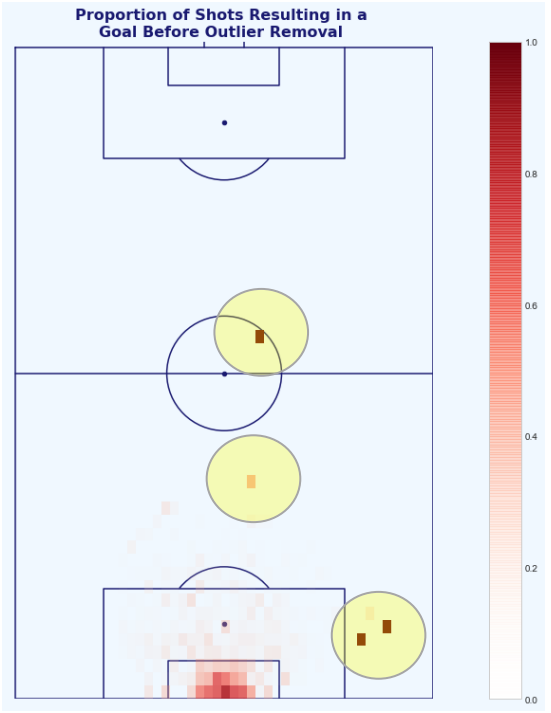


Fig. A: Heatmap of the proportions of shots to goals in the dataset (goals divided by shots). This visualisation flags some of the outlier shots scored from inside the attacking teams half, and at acute angles to the left of the keepers left post.

B

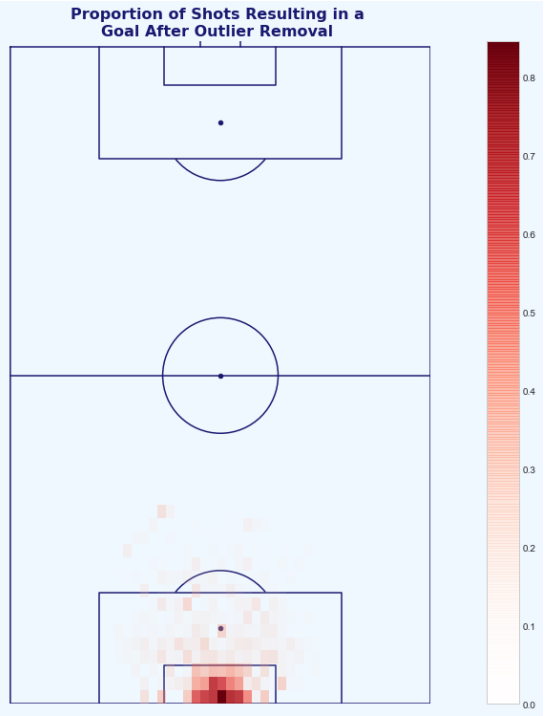


Fig. B: Heatmap of the proportions of shots to goals in the dataset post-outlier treatment. These outliers have now been removed to prevent this affecting the Chance Quality model.

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.
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.

Feature Engineering and Univariate Analysis

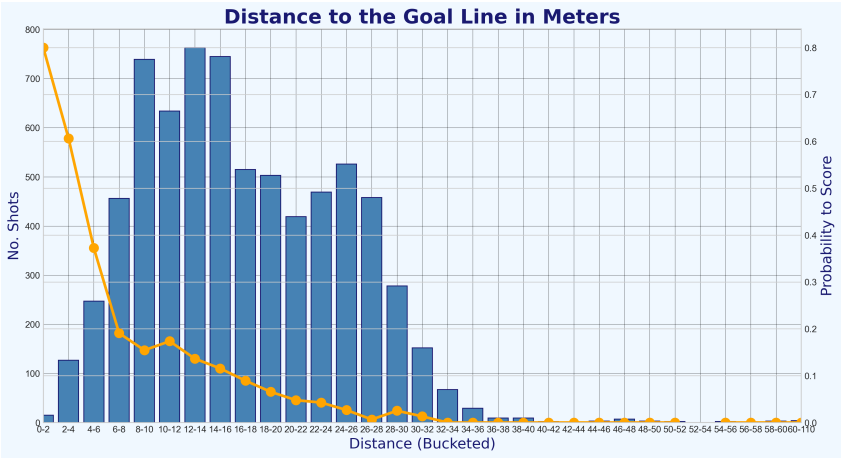
Creating and analysing the features used in the improved Chance Quality Model

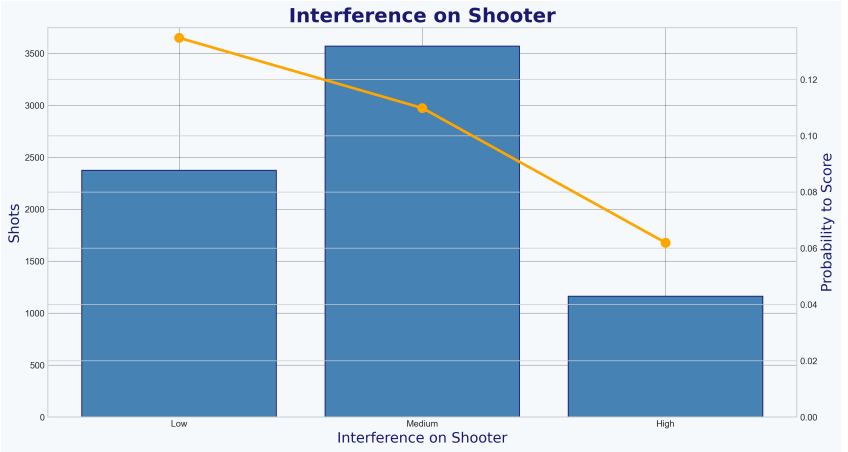
- **Feature Engineering considerations:**
  - Feature is available in both datasets;
  - The continuous and discrete features selected are monotonic in nature; and
  - Categorical features are required to be one-hot (dummy) encoded, with each value assigned it's own feature.

- **New features created for a Logistic Regression friendly model:**
  - 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; and
  - Dummy encoding of categorical features including Body Part to is_foot and is_head, and Interference on the Shooter to Low, Medium and High.

- **Monotonicity of the features:**
  - Figure A shows the monotonic relationship of distance of a shot and the probability to score.
  - Figure B shows that the interference level on the shooter, the higher the probability that they are to score from the shot (Fig. B).

- More information about the monotonicity of each feature can be found in Univariate Analysis (section 8) 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

Multivariate Analysis

Dealing with highly correlated features

- Considerations for correlation:
  - 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 figure on the right shows the correlation matrix of the final model.

- Highly correlated features such as Angle, Distance to the Center (M), Header were dropped, along with the Medium Interference on the Shooter (when one-hot encoding a categorical column, one column should be removed).

- 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 have assumed are not derivatives of each other and not correlated, and will therefore be left untouched in this exercise.

- More information about the Multivariate analysis and can be found in section 10 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

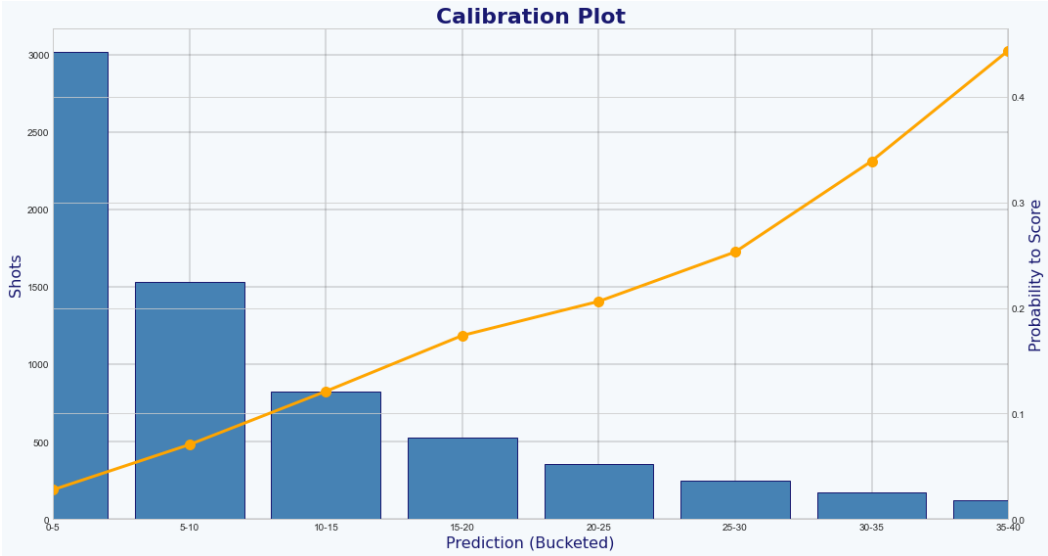
- **Final Logistic Regression model performance:**
  - Log Loss: 0.289 (reduction from 0.332)
  - ROC AUC: 80.7% (increase from 72.8%)

- **Model calibration:**
  - Calibration plots (Fig. A) are used to model probabilities correctly.
  - The observed likelihood of the shots in the dataset resulting in goals is between the 0% and 5% bucket. The model is therefore well calibrated.

- **Feature interpretation:**
  - Very important in football analytics as it is the intersection of Data Scientists working with Football Practitioners.
  - Finds in the model that leads to probabilities need to be able to be explained in a footballing context. This is done through the Feature Interpretation (Fig. B):
    - The further away the shot, the less likely the shot is to result in a goal;
    - The greater the number of intervening opponents, the less likely the shot is to result in a goal; and
    - The greater the header distance (i.e. the shot is a header), the less likely the shot is a goal.

- More information about the Final Model and Feature Interpretation can be found in sections 12 and 13 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 Metrica Sports Data

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

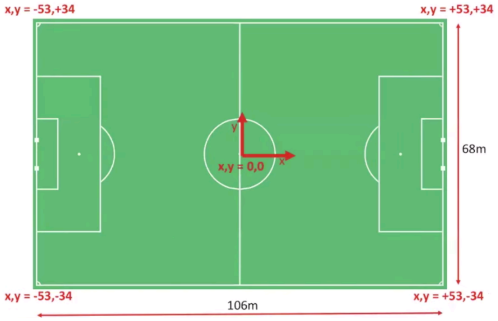
Metrica Sports Tracking and Event Data

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

1

Data Engineering

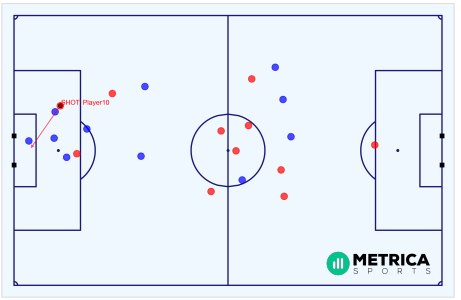
Conversion of the pitch coordinates, reversing the direction of players, determining each player's positions, speed, acceleration, and movement at a defined moment, and subsetting home and away DataFrames.



2

Exploratory Data Analysis of the match

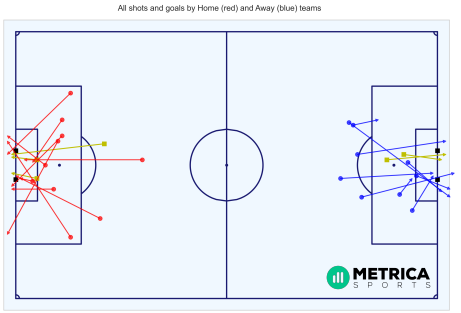
Analysis and visualisation of the twenty four shots that take place during the match by both the Home and Away teams.



3

Create features from Event and Tracking data required for CQM compatibility

Feature to add 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.



4

Export data

Finalised dataset of all shots with features ready for predictions and assessment of team performance using the CQM.

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

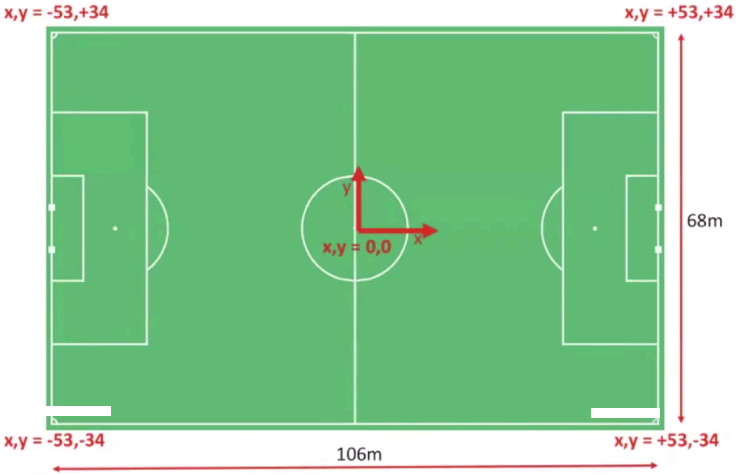
- **Conversion of Pitch Coordinates** (using the 'to_metric_coordinates' function from the Metrica IO (mio) library created by Laurie Shaw [link]):
  - Original: x(0, +1)m, y(0, +1)m
  - Converted: x(-53, +53)m, y(-34, +34)m (Fig. A)

- **Reverse player direction (same for the full 90 mins):** Home team right-to-left, Away team left-to-right.

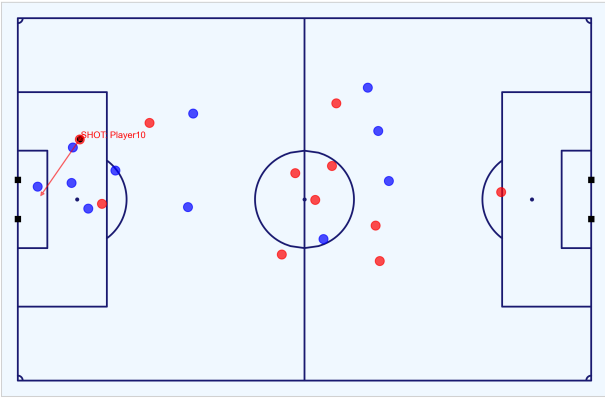
- **Determine each player's speed, acceleration and direction** (using the Metrica Velocity (mvel) library by Laurie Shaw [link], see Fig. B):
  - Metrica Sports Tracking data is collected at twenty five frames s^{-1}.
  - A player's speed can be calculated by dividing the distance a player has covered between two frames by 0.04 and using a Savitzky-Golay filter for smoothing.
  - The acceleration is calculated as the 2nd order derivative of speed.

- **Subset DataFrames:** Separate Home and Away DataFrames

A



B



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

Visualising the Home team's 3-2 win using the Event and Tracking data

- The match featured the following shots and goals:
  - Figure A: All the shots and goals for both the Home (red) and Away (blue teams). Goals are in yellow.
  - Figure B: First goal scored by the Home team at 8m8s (1-0).
  - Figure C: Second goal scored by the Away team at 35m22s (1-1).
  - Figure D : Third goal (header) scored by the Home team at 49m19s (2-1).
  - Figure E: Forth goal (penalty) scored by the Away team at 76m40s (2-2).
  - Figure F: 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.

- More information about each goal and the key chances can be found in the appendix or in the Metrica Sports notebook [link].

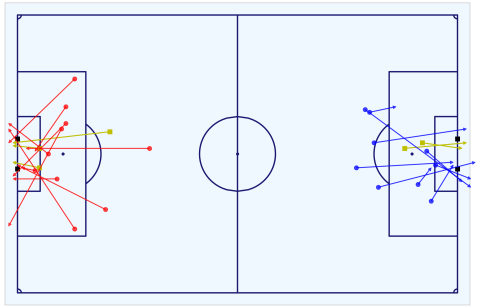


Fig. A: All shots and goals

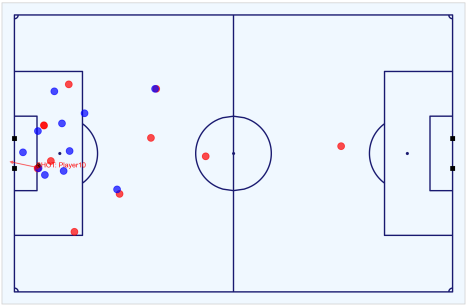


Fig. B: 1-0 Home goal (8m8s)

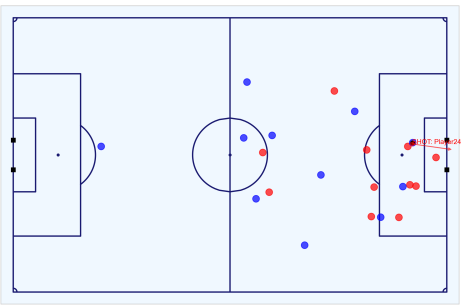


Fig. C: 1-1 Away goal (35m22s)

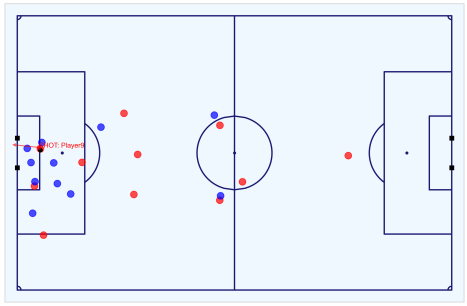


Fig. D: 2-1 Home goal (49m19s)

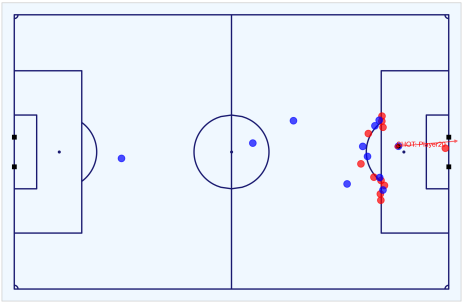


Fig. E: 2-2 Away penalty goal (76m40s)

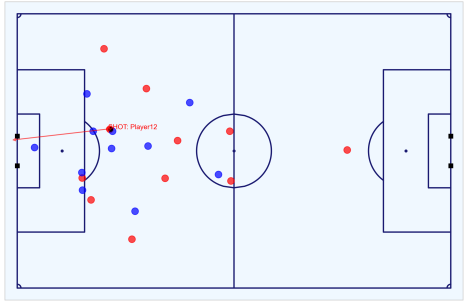


Fig. F: 3-2 Home goal (80m41s)

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 From Tracking Data

Features for compatibility with the CQM

- For compatibility with the trained CQM, a number of key features were required to be derived from the Event and Tracking data. These included
  - distance to the goal (m);
  - distance to the centre of the pitch (m);
  - angle to the goal (degrees);
  - 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, to be added to the Events data:
  - A players is determined as interfering if their distance between themselves and the ball is less than the defined radius of 5-yards (4.572m) - as used by StatsBomb to define pressure events [link][1].
  - A player (team mate or opponent) is determined as obstructing the goals from the viewpoint of the shooter (intervening) if they lie within the triangle between both posts and the ball

- The final Events DataFrame is the filtered for just the shots and then export for application with the trained CQM.

[1]How StatsBomb Data Helps Measure Counter-Pressing: https://statsbomb.com/2018/05/how-statsbomb-data-helps-measure-counter-pressing/
Sample Tracking data provided by Metrica Sports: https://github.com/metrica-sports/sample-data
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

A

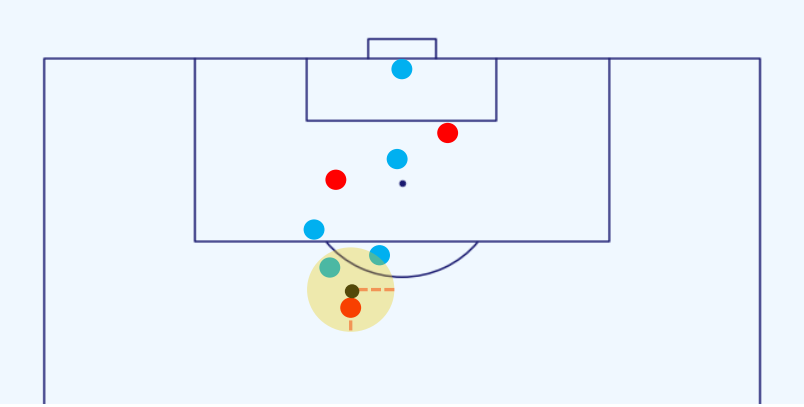


Fig. A: A player is defined as interfering with the shooter if they are within a defined radius of the ball, in this case 5m (not to scale). This can be determined mathematical by observing whether the distance between the player and the ball is less than the defined radius, or not.

B

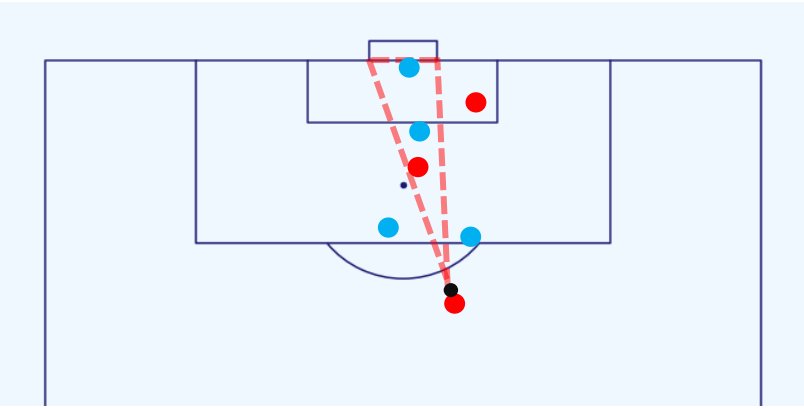


Fig. B: The number of intervening team mates and opponents can be visualised as a triangle between the ball and the posts of the goal at the moment the shot is taken. This interference can be determined mathematical by summating the areas of each triangle.

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

- Probabilities of the exported Metrica Sports data for the **Home team** in red and the **Away team** visualised as a cumulative xG race chart (see Fig. A).
- The match featured a penalty, however, the CQM was trained with just Open Play (OP) shots.
- Penalty kicks all share the same characteristics and were treated in this analysis by assigning values 0.76 xG. as per the value used by StatsBomb/FBref [link].
- What was a very tight contest regarding cumulative xG,1.43 to the Home side and 1.39 to the Away side (Fig. A), because a clearer projected win for the Away team, once assigned this new xG value for the penalty, 1.43 to the Home team and 2.00 to the Away team (Fig. B).
- The winning goal scored by the Home team in their 3-2 win was an 80th minute was a shot from distance with a low xG value of just 0.054 (ranked 13th out of 24).
- In this one-off game, the Away team was arguable are more deserving of winning the match by cumulative xG. However, the Home team capitalised on a low-xG chance in the last 10 minutes and went on to win a tight game.
- Shot-by-shot analysis of all the key chances can be found in the appendix.

A

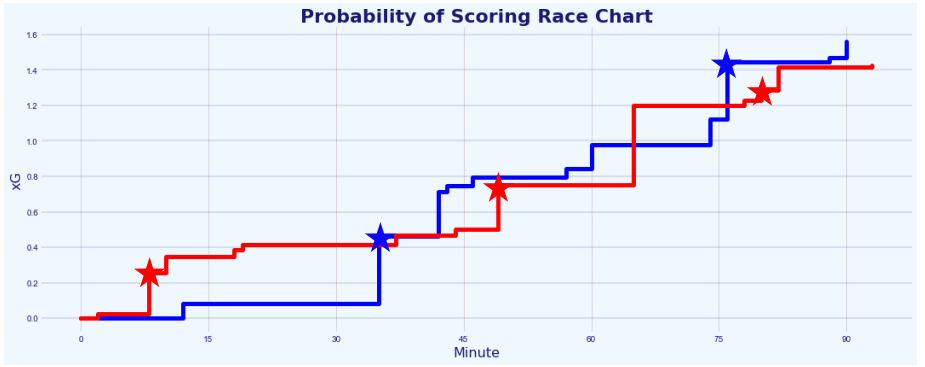


Fig. A: Probability of Scoring race chart before any amendments to xG values.

B

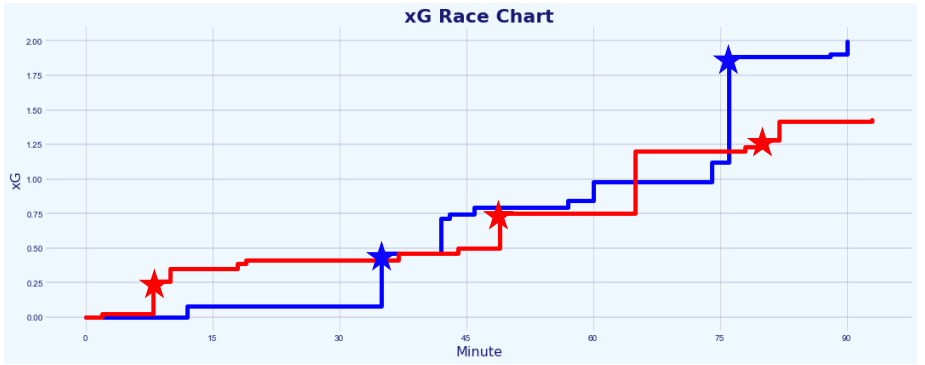


Fig. B: xG race chart with amended xG value for the penalty.

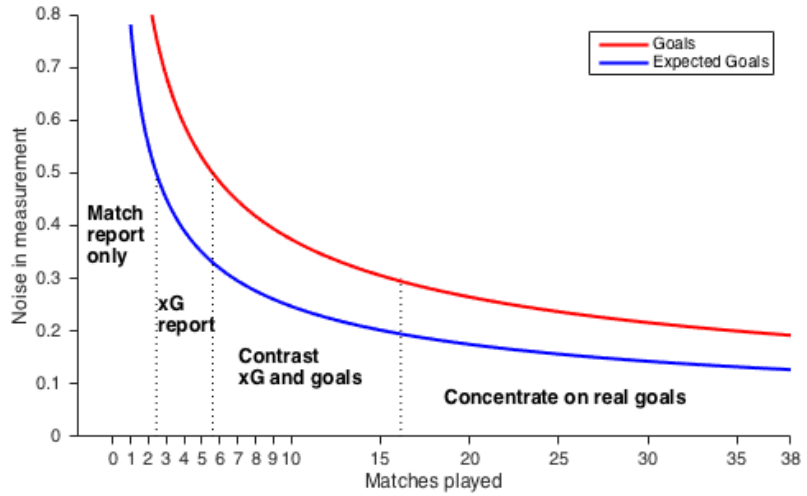
`xG explained by FBref (include penalty xG of 0.76): https://fbref.com/en/expected-goals-model-explained/
Analysis of game 2 of the Metrica Sports sample data can be found in section 14 and 15 of the Chance Quality Model notebook:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%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
Visualisations inspired by Ben Mayhew's Match Timelines, see: https://experimental361.com/explanations/match-timelines/
How to Create xG Flow Charts in Python by McKay Johns: https://www.youtube.com/watch?v=bvoOOYMQkac&list=PL10a1_q15HwqVEcnqt3tXs1bgvawjsQNW

With Great Power Comes Great Responsibility

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

- The concept of Expected Goals is a very useful concept because football is a low-scoring game. Subsequently, it has now entered the mainstream media lexicon[1].
- In his piece 'Should you write about real goals or expected goals?', David Sumpter demonstrates 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)[2].
- 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[2].
- The CQM suggests that the Away team is a slightly more deserving winner of the match by proxy of a greater accumulation of xG. However, 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 for 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

[1]See Laurie Shaw's article: Bodies on the Line: Quantifying how defenders affect chances: http://eightyfivepoints.blogspot.com/2017/09/bodies-on-line-quantifying-how.html.
[2]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.

Limitations of Chance Quality Models

Some of the known problems and issues with current modeling

- Any dataset analysed is incomplete, including the Shots dataset provided. With a full Event dataset, it is possible to include new features such as whether a goal was scored with the player's strongest foot, whether the shot was from a counter attack or fast break, whether the assist was a smart pass, whether the shot was from a cross, what the team's current game state is (winning/drawing/losing). Tracking data can then be used to add more features such as player and goalkeeper's position. However, despite being the state-of-the-art dataset, Tracking data is still missing significant datapoints including the body pose position (i.e. are they open to receive a pass) and the spin of the ball (therefore need to build in uncertainties when determining ball trajectory). More details about the advantages and disadvantages of using Event and Tracking data has been including in the appendix of this revised deck.

- Chance Quality models typically do not include information on the player taking the shot and are therefore an estimate of how the average player or team would perform in a similar situation[1]. This is because the model is built with all shot data, not just that individual players. For reference, currently in the 20/21 season after 30 matches, Harry Kane, has had 117 shots (45 on target, 21 goals), 3.9 shots a game. Over a 20 year career with 38 games per season, this would result in just under 3,000 shots.

- Chance Quality models also typically do not use data from after the point the player strikes the ball. This means the model does not include information after the shot, such as the direction, speed, and angle the ball is heading.

[1]FBref Expected Goals Explained: https://fbref.com/en/expected-goals-model-explained.
[2]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/ (using Wayback Machine).
[3]A new way to measure keepers shot stopping: post-shot expected goals by Mike Goodman: https://statsbomb.com/2018/11/a-new-way-to-measure-keepers-shot-stopping-post-shot-expected-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

Comparison of Logistic Regression Model with XGBoost Model

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

- Due to time constraints and because model performance was not the absolute priority of this data challenge, the CQM submitted was trained using just a Logistic Regression model.
- A subsequent Gradient Boosting Algorithm model has been created using XGBoost (eXtreme Gradient Boosting), that has a slightly improved performance in terms of Log Loss compared to the initial model.
- XGBoost is based on decision trees and has a great track record of high performances on modeling structured data due to its performance, flexibility and speed, and is regularly the algorithm that wins Kaggle competitions.
- Much more information about how this second model was trained using XGBoost can be found in the separate Gradient Boosted Chance Quality Model notebook [link].

A

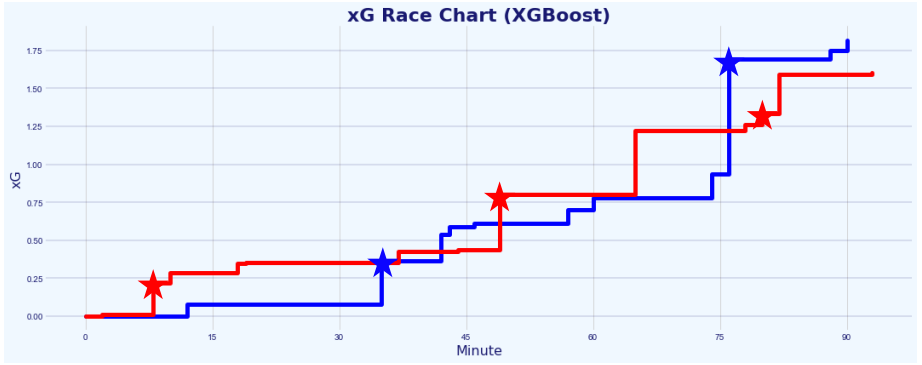


Fig. A: xG race chart for Chance Quality Model created using XGBoost.

B

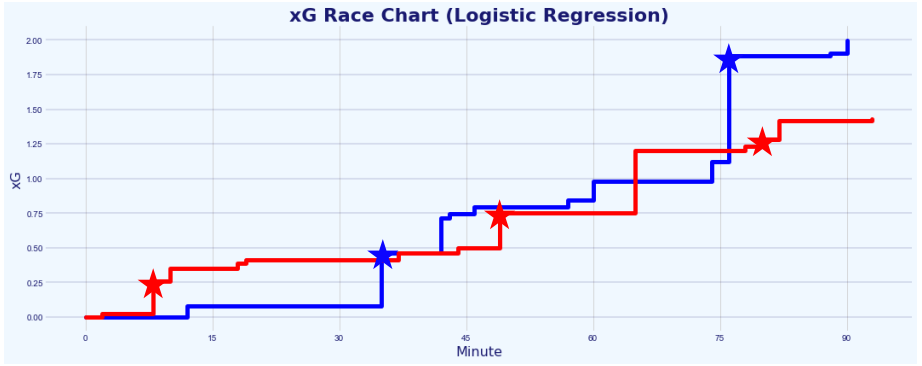


Fig. A: xG race chart for Chance Quality Model created using Logistic Regression.

Secondary model created using XGBoost in a separate notebook:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/2)%20Gradient%20Boosted%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb
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

Next Steps and Conclusion

Conclusion

Summary of this data challenge submission

1. Defined, built, trained, and evaluated two Chance Quality Models from Shots data through the application of both Logistic Regression and XGBoost, where Expected Goals was the KPI used to interpret model predictions and assess shot quality.

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 Models in an appropriate format, deriving additional features from the Tracking data to further enrich the provided Event data, to be used in the Chance Quality Model predictions.

3. Assessed the performance of the two 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
      - Chance Quality Model notebook (Logistic Regression):
      - Chance Quality Model notebook (XGBoost):
    - 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 pack (this one): https://docs.google.com/presentation/d/116D0U_ue2sv6hgLBgHnil228cRhph0qfMuOFA4uuxhs/edit?usp=sharing
  - Data Science pack (as part of the 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, focussing on Tableau): 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

Further Work

eddwebster/
football_analytics

⚽📊 A collection of football analytics projects, data, and analysis by Edd Webster (@eddwebster), with links to publicly available resource...

1 Contributors | 0 Issues | 94 Stars | 6 Forks

Additional Projects Regarding Football Analysis 1/3

Democratisation of data using Tableau to build tools for football practitioners at all levels of the club

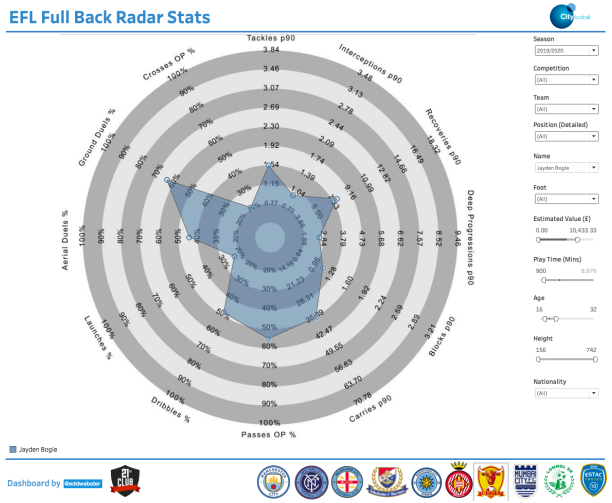
- Dashboards available @ public.tableau.com/profile/edd.webster.
  - WSL dashboards and analysis [link] ;
  - ‘Big 5’ European leagues dashboards and analysis [link];
  - EFL dashboards and analysis [link];
  - StrataBet European Leagues Chance analysis [link]; and
  - Opta #mcfcanalytics dashboards and analysis [link].

- Notable examples of reporting tools include:
  - Position-specific Radar dashboards to visualise a large number of player statistics at one time (Fig. A);
  - Passing Network analysis to observe the ball exchange between players (Fig. B);
  - Formation analysis based on most historically used formations and average position of players (Fig. C); and
  - Individual match xG visualisations to see the chances created and most impactful players through an xG race chart and on-pitch shooting positions (Fig. D).

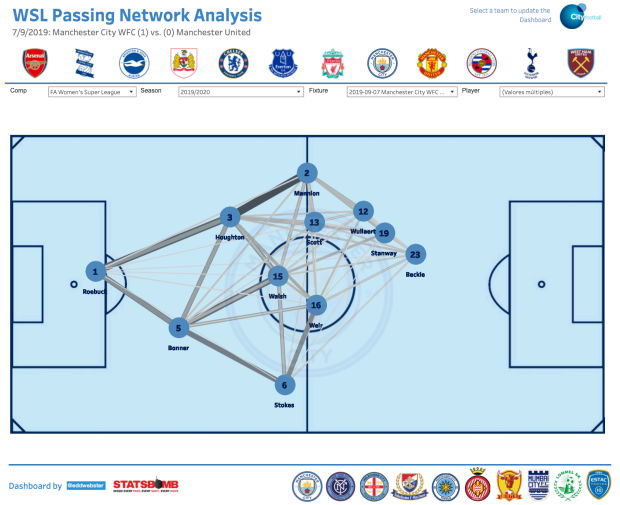
- To see the full portfolio of data reporting tools, please see the following link for a full PowerPoint presentation in which each tool and the build process is explained in detail: https://docs.google.com/presentation/d/1uT6vT1J_FS10hW3hvWQx9ofqHUq8mT-kUbDRMrrUqH8/edit?usp=sharing.

- For the notebooks that scrape, parse, engineer and unify the datasets used in these visualisations, see: github.com/eddwebster/football_analytics/tree/master/notebooks.

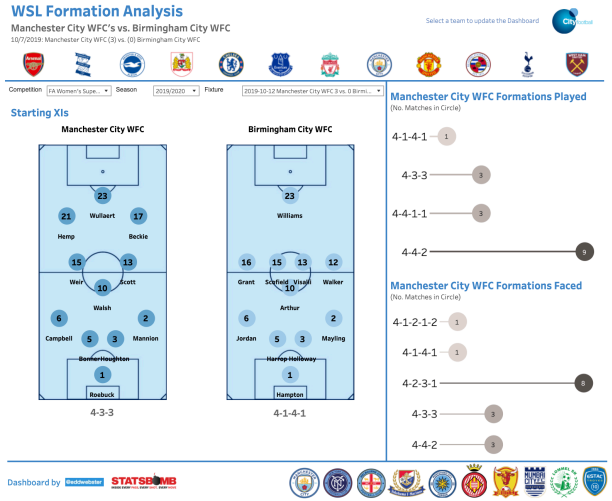
A



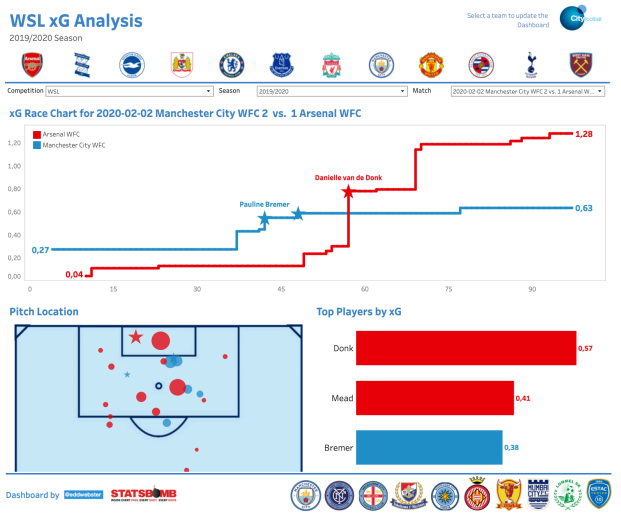
B



C

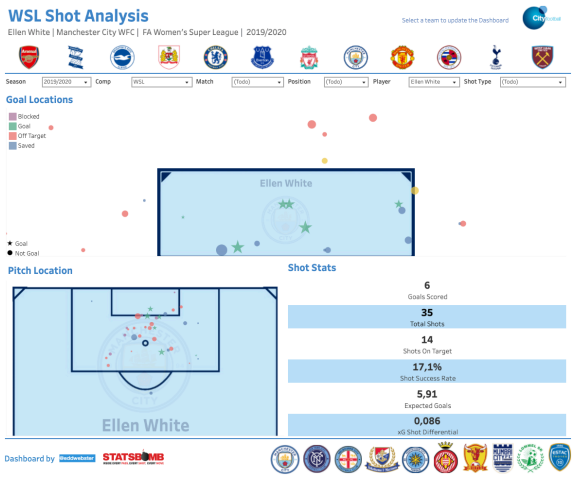


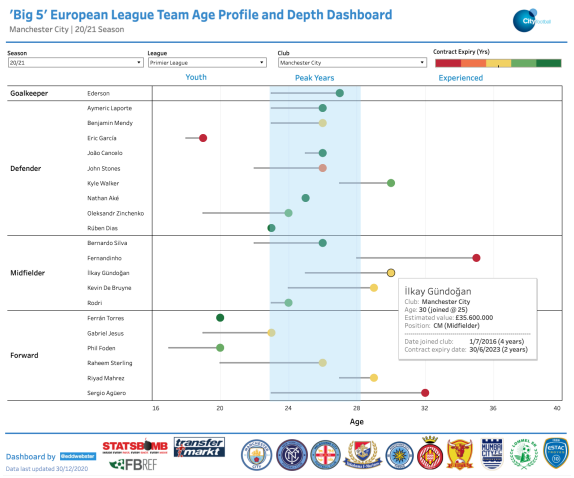
D

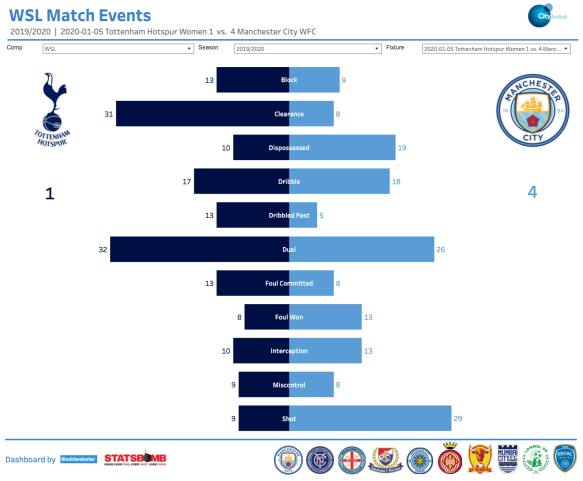


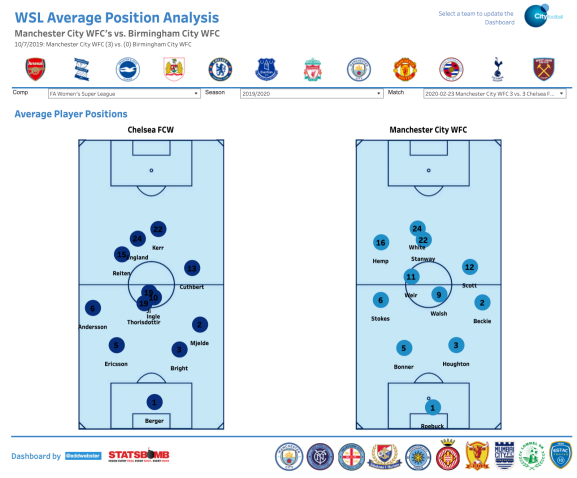
Additional Projects Regarding Football Analysis 2/3

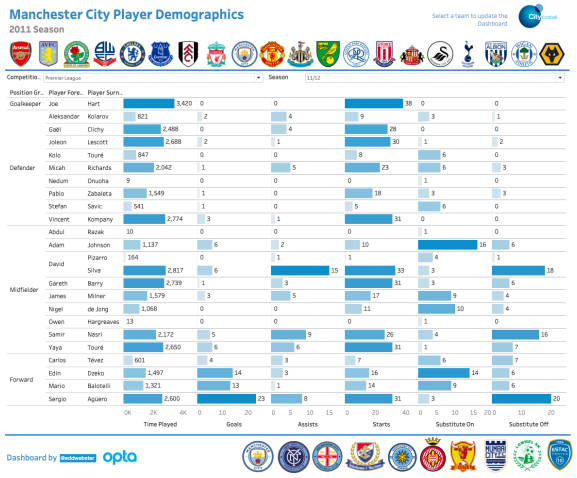
More Tableau reporting tools

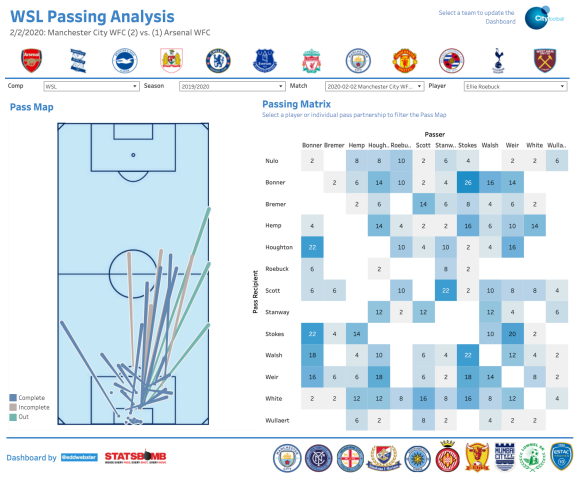


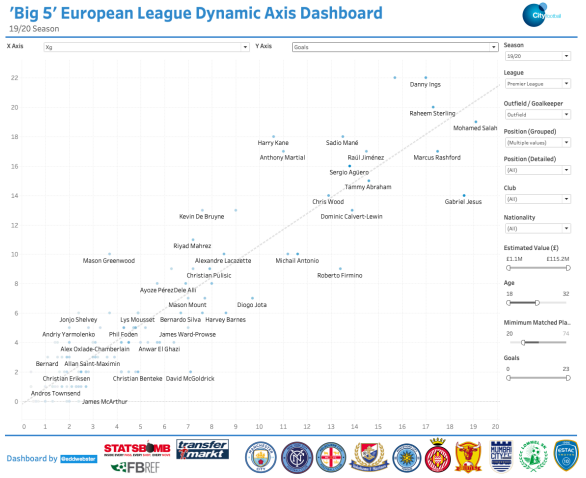


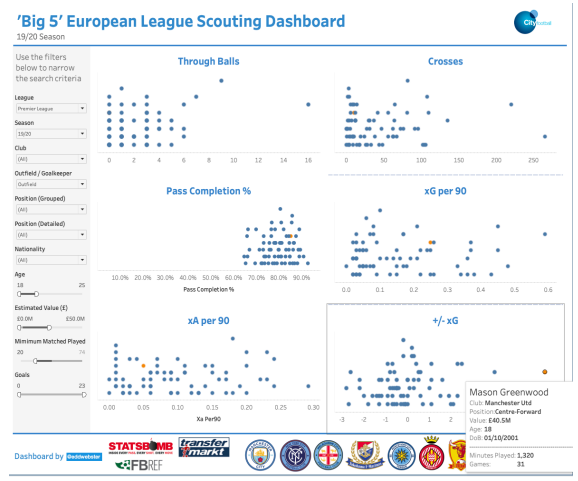








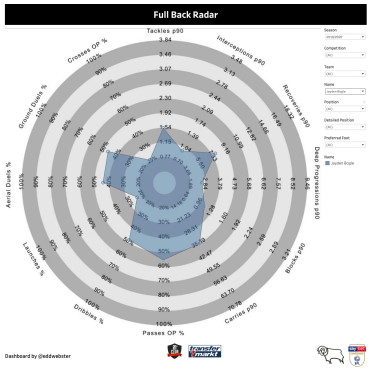




Additional Projects Regarding Football Analysis 3/3

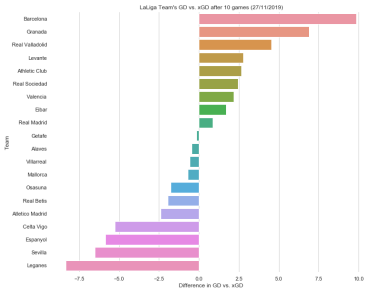
Links to additional projects that demonstrate a broad range of tools, skills and analysis when working with football data

- **Derby County recruitment analysis**
  - This project uses a data-driven approach to provide a number of specified recommendations for replacements of Derby County's right-back Jayden Bogle and right-winger Tom Lawrence. This project uses aggregated player performance data provided by 21st Club matched against estimated player valuation data scraped from TransferMarkt
  - Slides: https://docs.google.com/presentation/d/1qLdXpcO7tYPIojHmDXny3hFAPw1h7q2-FbNCZuq46ls/edit?usp=sharing
  - Tableau: https://public.tableau.com/views/EddWebsterDerbyCountyDataAnalysis/RBDemographicsDashboard?:language=es&:display_count=y&:origin=viz_share_link
  - Main notebook (several can be found in the GitHub repo): https://github.com/eddwebster/derby_county_data_task/blob/master/notebooks/B)%20Data%20Engineering/Data%20Engineering%20of%2021st%20Club%20and%20TransferMarkt%20Datasets.ipynb
  - GitHub: https://github.com/eddwebster/derby_county_data_task/



- **FIFA-TransferMarkt fantasy football league**
  - This project scrapes data for 40+ leagues on TransferMarkt using Beautifulsoup and matches it to the 18,000 players in the FIFA 20 database, using the record linkage library, as part of a data-driven approach of £/attribute metrics for a hypothetical recruitment scenario. This data is then exported to Excel as part of a playable fantasy football league.
  - Notebook: https://nbviewer.jupyter.org/github/eddwebster/fifa-league/blob/master/FIFA%2020%20Fantasy%20Football%20League%20using%20TransferMarkt%20Player%20Valuations.ipynb
  - Functioning Microsoft Excel league spreadsheet for player draft and match result data entry: https://github.com/eddwebster/fifa-league/blob/master/excel/fifa_20_fantasy_football_league.xlsm
  - GitHub: https://github.com/eddwebster/fifa-league

- **Granada 19/20 xG analysis**
  - Exploratory Data Analysis (EDA) of scraped xG data from the first ten games of the 19/20 season of La Liga, analysing Granada's overachieving performance after promotion to Spain's top flight.
  - Notebook: https://nbviewer.jupyter.org/github/eddwebster/granada1920/blob/master/Granada%202019-20%20Web%20Scraping%20and%20pandas%20-%2012012020.ipynb
  - GitHub: https://github.com/eddwebster/granada1920



References

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

Chance Quality/Expected Goals modeling

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/
- Bodies on the Line: Quantifying how defenders affect chances by Laurie Shaw: http://eightyfivepoints.blogspot.com/2017/09/bodies-on-line-quantifying-how.html.
- Should you write about real goals or expected goals? A guide for journalists by David Sumpter: 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/
- Match Timelines by Ben Mayhew: https://experimental361.com/explanations/match-timelines/
- 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

Appendix

Key Chances Analysis 1/8 – Highest Quality Chances and All Goals Scored

Isolating the seven shots that make up the the five shots with the highest probability and the two goals outside that

- As per the submitted, Logistic Regression model, the five best chances in the game in terms of xG are the following:
  1) Shot 18 - penalty to the Away team with xG of 0.760 (goal)
  2) Shot 16 - shot by the Home team with xG of 0.450 (saved)
  3) Shot 7 - shot by the Away team with xG of 0.379 (goal)
  4) Shot 9 – headed shot by the Away team with xG of 0.232 (missed)
  5) Shot 13 – headed shot by the Home team with xG of 0.251 (goal)

- Two of the goals scored were not in the top five chances in terms of xG:
  - Shot 2 - shot by the Home team with xG of 0.235 – ranked 6th for xG
  - Shot 20 - shot by the Home team with xG of 0.054 – ranked 13th for xG

- This selection of shots includes all the chances with a 20% or greater chance of resulting in a goal.

- The following slides go through each of these seven chances, one-by-one, used as part of the analysis to assess which team was the more deserving to win the match, when considering xG for a single game.

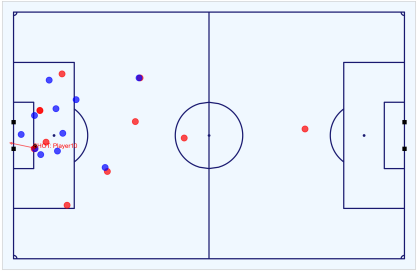


Fig. A: 8m8s - Shot 2 (Home) - goal

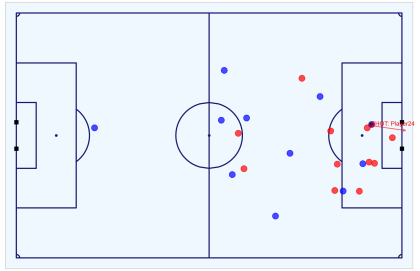


Fig. B: 35m22s - Shot 7 (Away) - goal

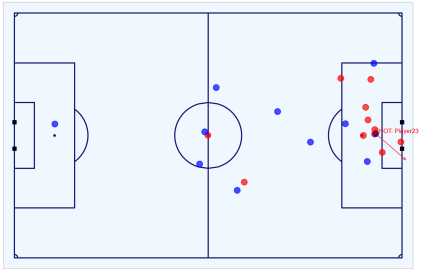


Fig. C: 37m39s - Shot 9 (Away) – no goal

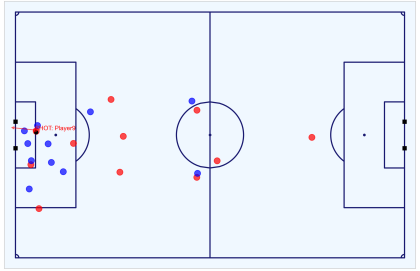


Fig. D: 49m19s - Shot 13 (Home) - goal

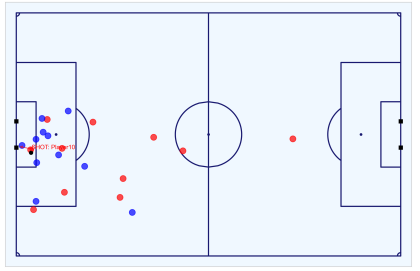


Fig. E: 65m93s - Shot 16 (Away) – no goal

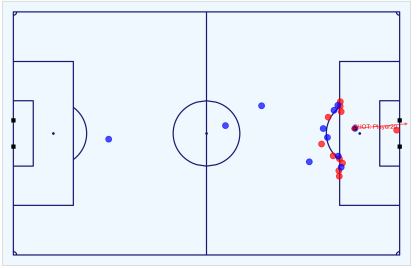


Fig. F: 76m40s - Shot 18 (Away) - goal

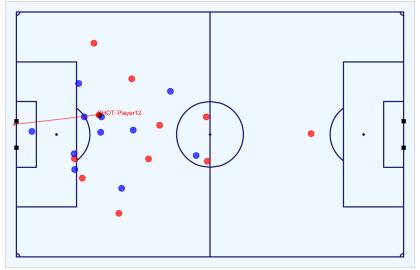


Fig. G: 80m41s - Shot 20 (Home) – goal

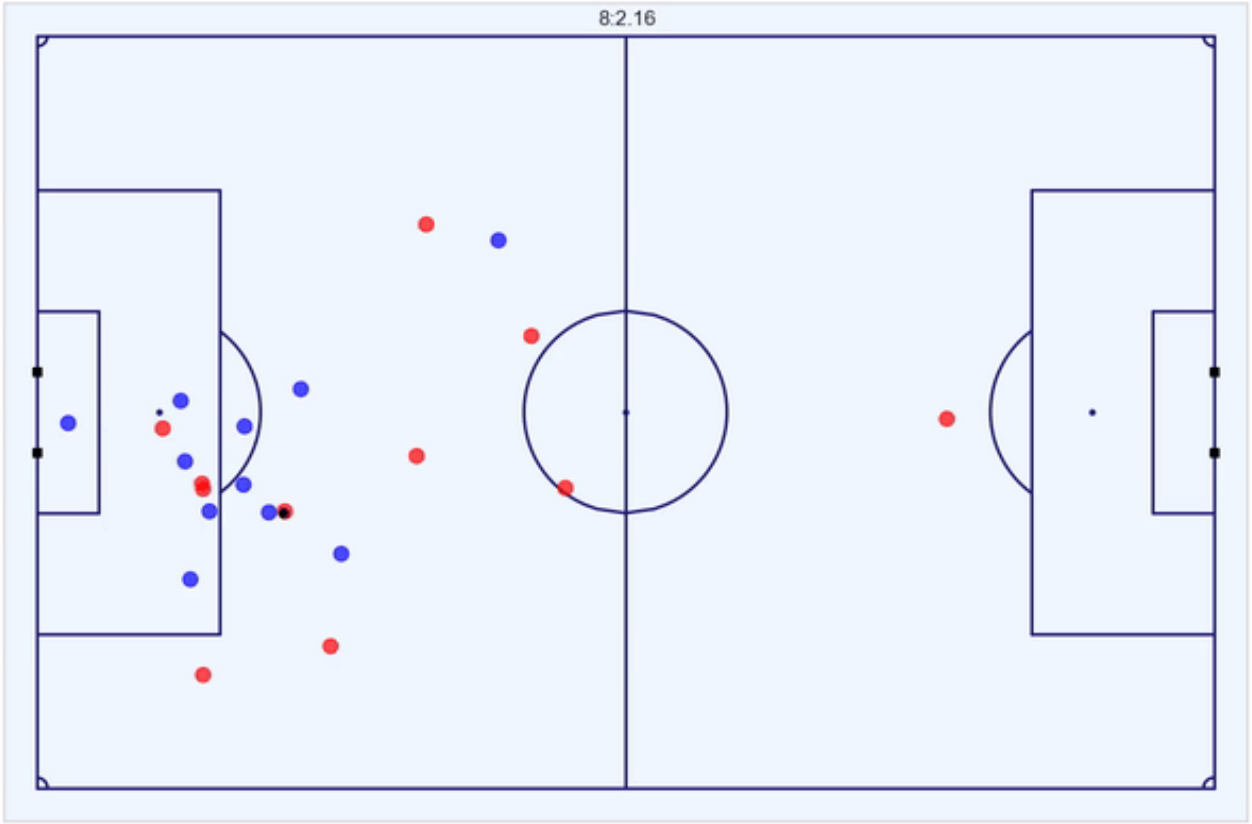
All GIF figures produced from Tracking data: https://github.com/eddwebster/mcfc_submission/tree/main/gif/fig/metrica-sports.
All MP4 figures produced from Tracking data: https://github.com/eddwebster/mcfc_submission/tree/main/video/fig/metrica-sports.
Notebook to work with the Metrica Sports Event and 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.

Key Chances Analysis 2/8 - Goal 1 of 5 (Shot 2)

Still of shot and the player positions for the first of the five goals scored

- **Shot details:**
  - Shot no. (of 24): 2
  - Time: 8m8s
  - Index (Frame): 198 (12,202)
  - Team: Home
  - Shot-taker: Player 10
  - Head or foot: Foot
  - Number of intervening opponents: 1
  - Number of intervening teammates: 0
  - Interference on shooter (players): High (2)
  - Subtype: On target-goal
  - Outcome: Goal
  - Penalty?: No
  - Direct Freekick: No
  - xG (LR): 0.235 (rank 6 of 24)
  - xG (XGB): 0.210 (rank 6 of 24)
- **Video:** [link]
- **Analysis:**

  First goal to the Home team comes from a cross from outside the box on the right hand side that cuts out all the defending players, to leave a tap-in from 6-yards-out with only the goalkeeper between the ball and the goal.

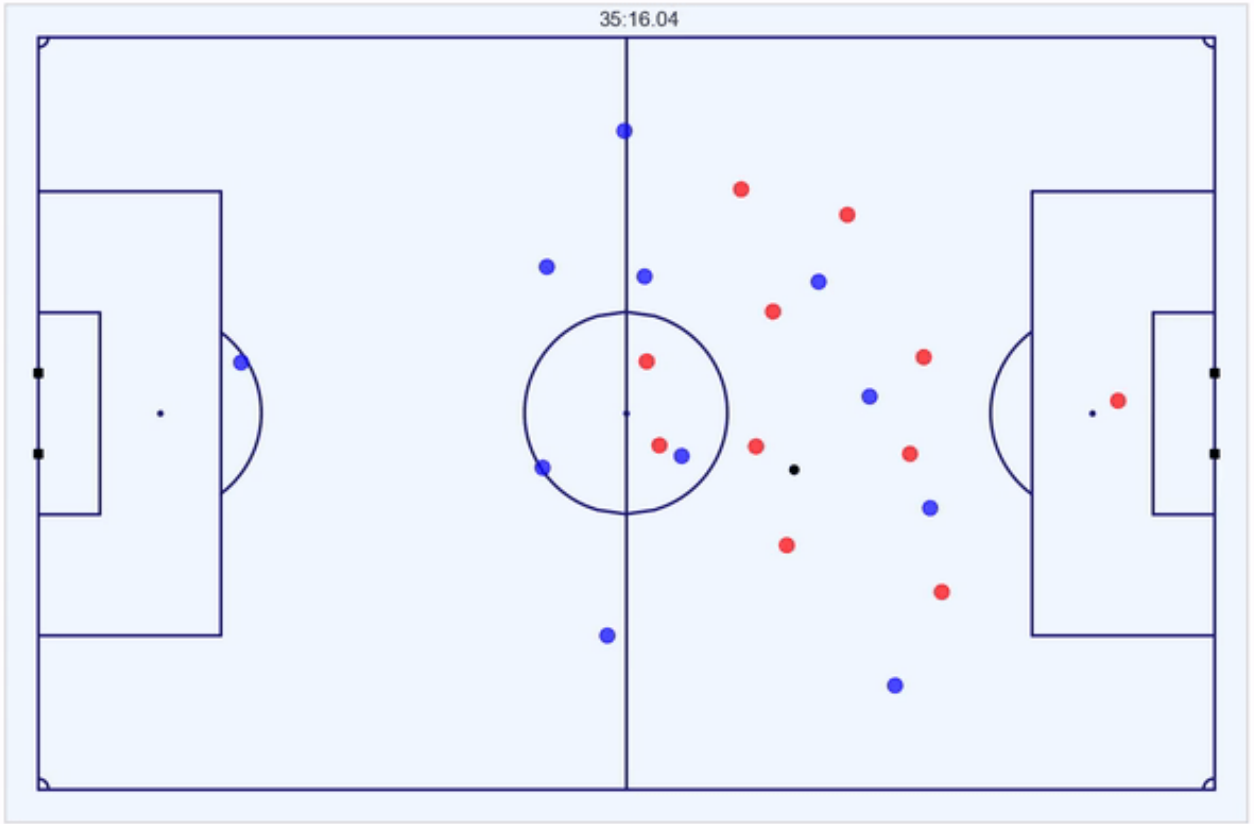


Notebook that analyses the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Notebook to create the Chance Quality Model from the sample shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.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.

Key Chances Analysis 3/8 - Goal 2 of 5 (Shot 7)

Still of shot and the player positions for the second of the five goals scored

- **Shot details:**
  - Shot no. (of 24): 7
  - Time: 35m22s
  - Index (Frame): 823 (53,049)
  - Team: Away
  - Shot-taker: Player 24
  - Head or foot: Foot
  - Number of intervening opponents: 1
  - Number of intervening teammates: 0
  - Interference on shooter (players): Low (1)
  - Subtype: On target-goal
  - Outcome: Goal
  - Penalty?: No
  - Direct Freekick: No
  - xG (LR): 0.379 (rank 3 of 24)
  - xG (XGB): 0.282 (rank 4 of 24)



- **Video:** [link]

- **Analysis:**

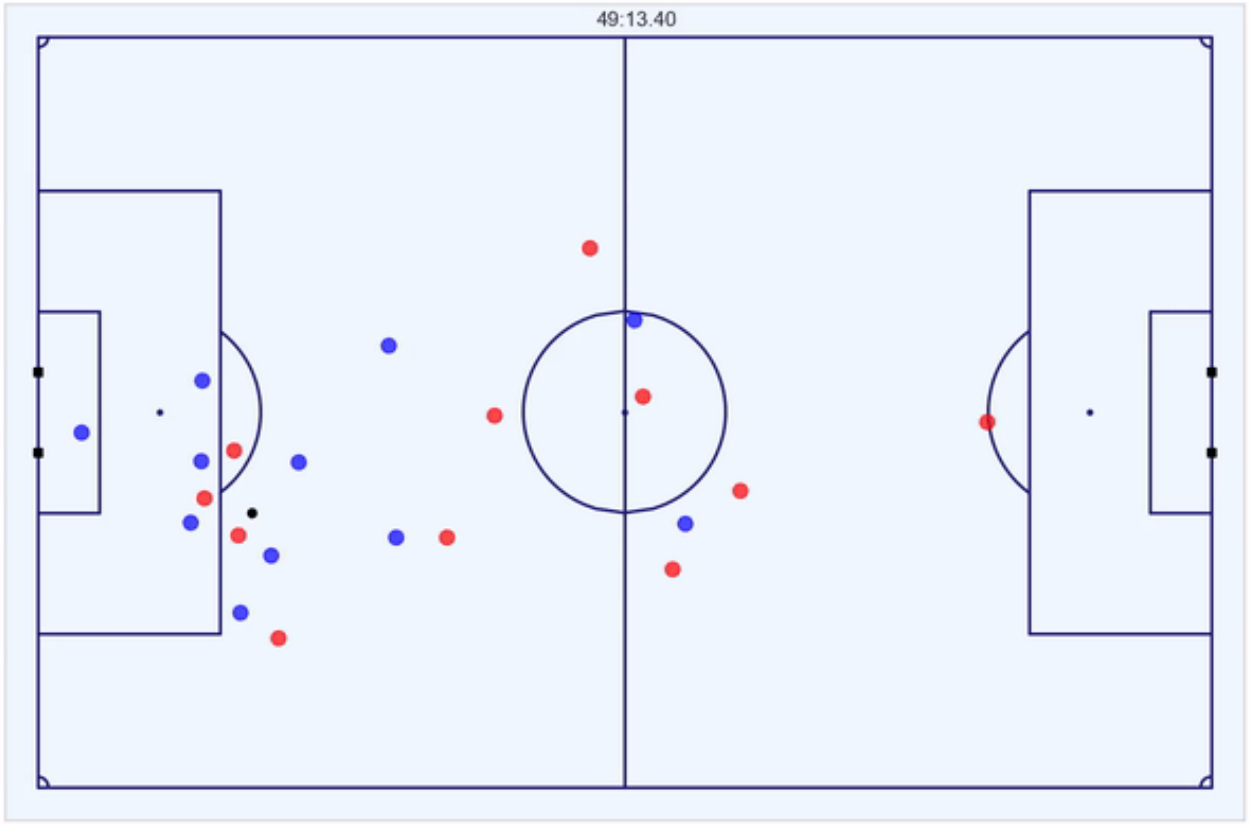
  First goal to the Away team and the 2nd goal in the match comes from a cross around 12-yards-out inside the right-hand side of the box cuts out all the defending players, to leave a tap-in from 8-9-yards-out with only the goalkeeper between the ball and the goal.

Notebook that analyses the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Notebook to create the Chance Quality Model from the sample shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.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.

Key Chances Analysis 4/8 - Goal 3 of 5 (Shot 13)

Still of shot and the player positions for the third of the five goals scored

- **Shot details:**
  - Shot no. (of 24): 13
  - Time: 49m19s
  - Index (Frame): 1,118 (73,983)
  - Team: Home
  - Shot-taker: Player 9
  - Head or foot: Head
  - Number of intervening opponents: 2
  - Number of intervening teammates: 0
  - Interference on shooter (players): Medium (2)
  - Subtype: On target-goal
  - Outcome: Goal
  - Penalty?: No
  - Direct Freekick: No
  - xG (LR): 0.251 (rank 5 of 24)
  - xG (XGB): 0.363 (rank 3 of 24)



- **Video:** [link]

- **Analysis:**

  Second goal to the Away team and the 3rd goal in the match comes from a cross around 12-yards-out just inside the left-hand side of the box that cuts out all the defending players, to leave a tap-in from 6 yards out with only the goalkeeper and a defender on the near post between the ball and the goal.

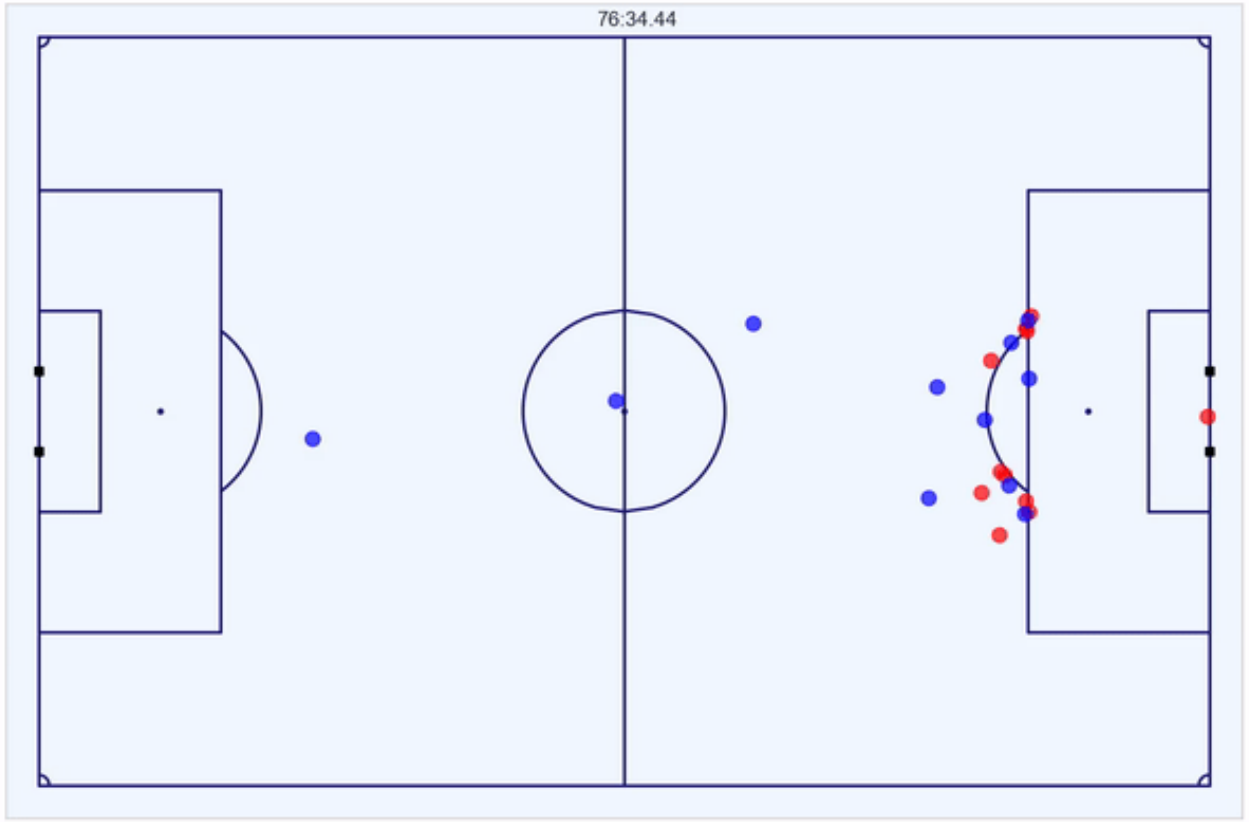
Notebook that analyses the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Notebook to create the Chance Quality Model from the sample shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.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.

Key Chances Analysis 5/8 - Goal 4 of 5 (Shot 18)

Still of shot and the player positions for the forth of the five goals scored

- **Shot details:**
  - Shot no. (of 24): 18
  - Time: 76m40s
  - Index (Frame): 1,671 (115,009)
  - Team: Away
  - Shot-taker: Player 20
  - Head or foot: Foot
  - Number of intervening opponents: 1
  - Number of intervening teammates: 0
  - Interference on shooter (players): Low (0)
  - Subtype: On target-goal
  - Outcome: Goal
  - Penalty?: Yes
  - Direct Freekick: No
  - xG (LR): 0.760 (as defined by StatsBomb) (rank 1 of 24)
  - xG (XGB): 0.760 (as defined by StatsBomb) (rank 1 of 24)
- **Video:** [link]
- **Analysis:**

  Second goal to the Away team and the 4th goal in the match comes from a penalty.

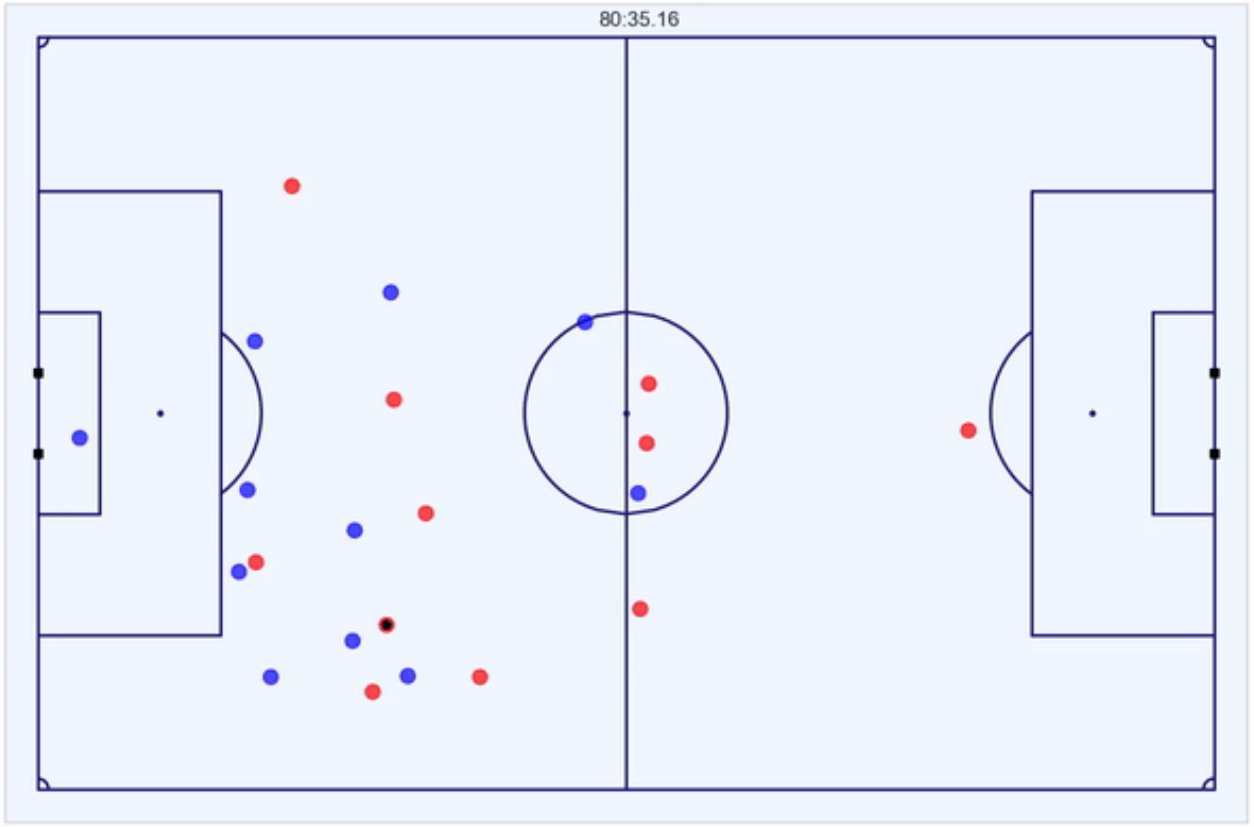


Notebook that analyses the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Notebook to create the Chance Quality Model from the sample shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.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.

Key Chances Analysis 6/8 - Goal 5 of 5 (Shot 20)

Still of shot and the player positions for the fifth of the five goals scored

- **Shot details:**
  - Shot no. (of 24): 20
  - Time: 80m41s
  - Index (Frame): 1,723 (121,027)
  - Team: Home
  - Shot-taker: Player 12
  - Head or foot: Foot
  - Number of intervening opponents: 2
  - Number of intervening teammates: 0
  - Interference on shooter (players): High (2)
  - Subtype: On target-goal
  - Outcome: Goal
  - Penalty?: No
  - Direct Freekick: No
  - xG (LR): 0.054 (rank 13 of 24)
  - xG (XGB): 0.074 (rank 13 of 24)



- **Video:** [link]

- **Analysis:**

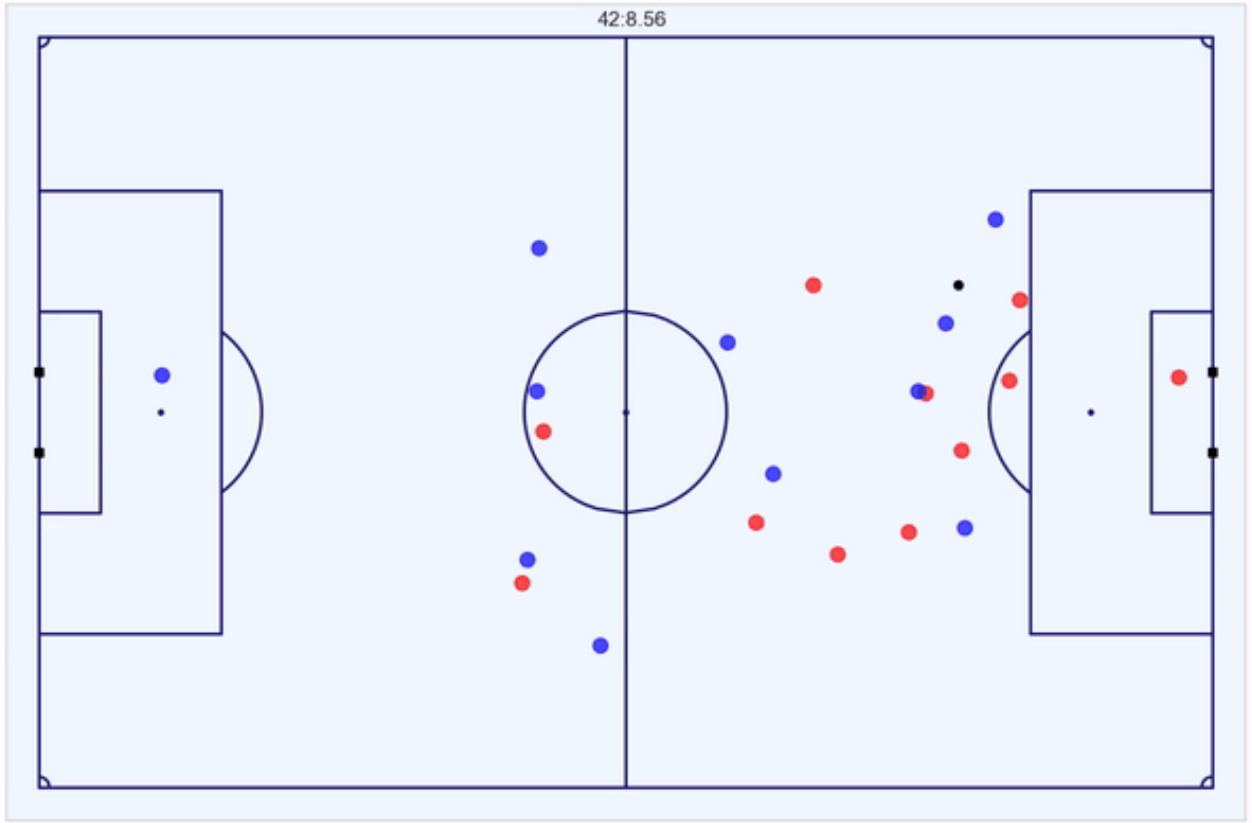
  Third goal to the Away team and the 5th and final goal in the match comes from a long shot from around 30-yards out, with two players, the goalkeeper and a defender obstructing the goal and two defending players pressuring the shot taker.

Notebook that analyses the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Notebook to create the Chance Quality Model from the sample shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.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.

Key Chances Analysis 7/8 - Shot 9

Still of shot and the player positions for the 9th shot (no goal)

- **Shot details:**
  - Shot no. (of 24): 9
  - Time: 37m39s
  - Index (Frame): 962 (63,362)
  - Team: Away
  - Shot-taker: Player 23
  - Head or foot: Foot
  - Number of intervening opponents: 1
  - Number of intervening teammates: 0
  - Interference on shooter (players): Medium (1)
  - Subtype: Off target-head-out
  - Outcome: No-goal
  - Penalty?: No
  - Direct Freekick: No
  - xG (LR): 0.450 (rank 4 of 24)
  - xG (XGB): 0.419 (rank 7 of 24)



- **Video:** [link]

- **Analysis:**

  Ninth shot in the match was to the Away team, coming from a cross from around 12-yards-out just inside the left-hand side of the box that cuts out all the defending players, to leave a tap-in from 7/8 yards with only the goalkeeper to beat, that was subsequently put wide.

Notebook that analyses the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Notebook to create the Chance Quality Model from the sample shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.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.

Key Chances Analysis 8/8 - Shot 16

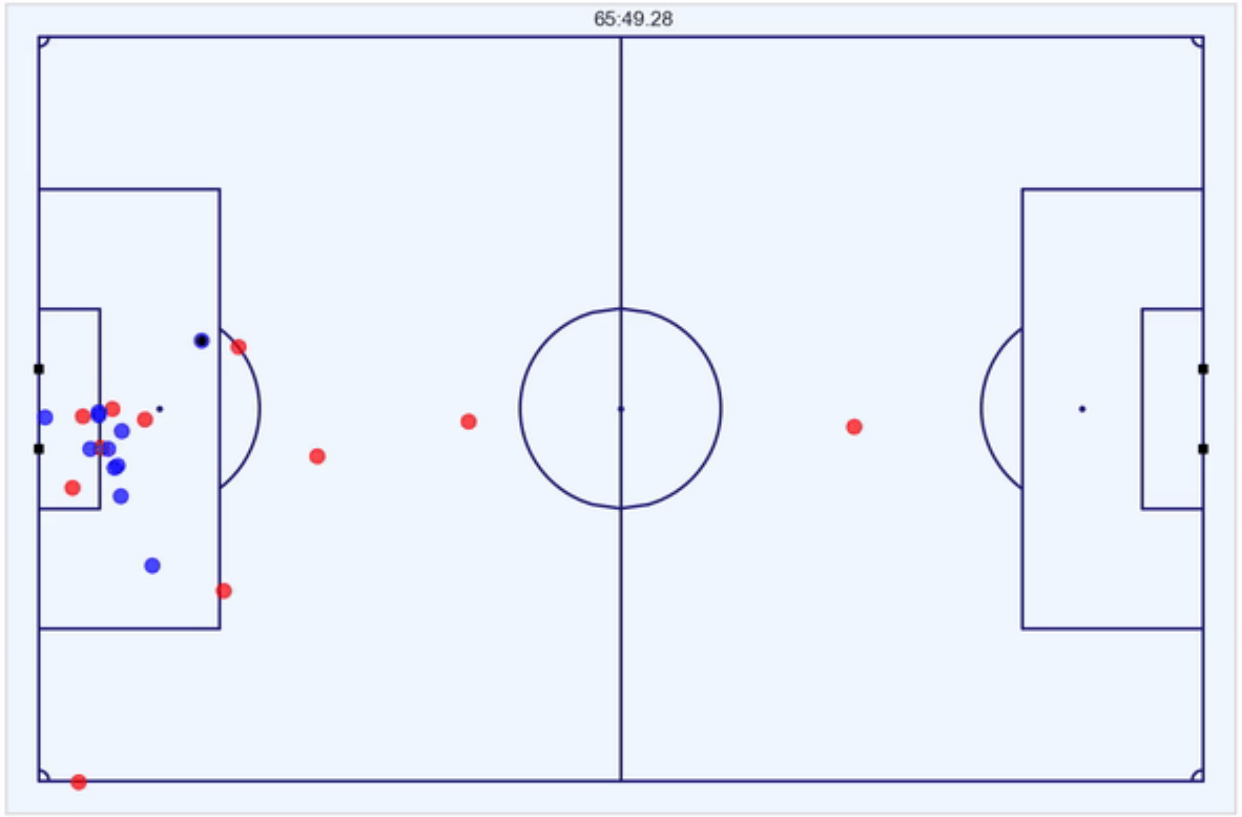
Still of shot and the player positions for the 16th shot (no goal)

- **Shot details:**
  - Shot no. (of 24): 16
  - Time: 65m93s
  - Index (Frame): 1,479 (98,880)
  - Team: Home
  - Shot-taker: Player 10
  - Head or foot: Head
  - Number of intervening opponents: 1
  - Number of intervening teammates: 0
  - Interference on shooter (players): Low (0)
  - Subtype: On target-saved
  - Outcome: No-goal
  - Penalty?: No
  - Direct Freekick: No
  - xG (LR): 0.252 (rank 2 of 24)
  - xG (XGB): 0.176 (rank 2 of 24)

- **Video:** [link]

- **Analysis:**

  Sixteenth shot in the match was to the Home team, coming from a cross from outside the box to the left-hand side, that is met with a header but saved by the goalkeeper.



Notebook that analyses the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Notebook to create the Chance Quality Model from the sample shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.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.

Model Selection 1/4: Algorithms Used - Logistic Regression...and then XGBoost

Background information about how Logistic Regression works

- Creating a Chance Quality model is a classic classic supervised learning problem. A classification model is tasked with drawing some conclusion from the input values. It will predict the class labels/categories for the new data. In the case of a Chance Quality model, whether a shot results in a goal or not.
- The initial model model (submitted for initial task deadline) was a Logistic Regression model. A subsequent Gradient Boosted Decision Trees model, using the XGBoost algorithm was used as a secondary model (see next slide and the separate Gradient Boosted Chance Quality Model notebook [link] for more information)[1].
- Logistic regression is a simple linear classifier that takes a weighted combination of the input features, and passes it through a function such as a sigmoid function[2], that maps any real number to a number between 0 and 1. Mathematically the equation for the sigmoid function can be represented as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

where e is Euler's number.

- The mathematical representation of the Logistic Regression function can be represented as:

$$f_{\mathbf{w},b}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{1 + e^{-(\mathbf{w}\mathbf{x}+b)}}$$

where:
- $\mathbf{w}$ is a D-dimensional vector of parameters
- b is bias (a real number)
- $f_{\mathbf{w},b}$ means that the model is parametrised by two values, $\mathbf{w}$ and b

- A Logistic classifier generally predicts the positive class if the sigmoid output is greater than 0.5, and the negative class otherwise. The Machine Learning problem is to find a weight combination that minimises the error, which in the case of Logistic Regression is the Logistic Loss.
- The optimisation criterion in Logistic Regression is called the maximum likelihood. Instead of minimising the average loss like in Linear Regression, Logistic Regression looks to maximise the likelihood of the training data according to the model.

[1]Logistic Regression wiki: https://en.wikipedia.org/wiki/Logistic_regression.
[2]Sigmoid function wiki: https://en.wikipedia.org/wiki/Sigmoid_function.
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

Model Selection 2/4: Algorithms Used - XGBoost

Background information about how XGBoost works

- The alternative to a Logistic Regression model is to use a Decision Tree-based method. Before defining Gradient Boosted Trees, it is first necessary to define a Decision Tree. A Decision Tree is an acyclic graph that can be used to make decision. Decision trees essentially learn a hierarchy of if/else question, leading to a decision. In each branching node of the graph, a specific feature, of the feature vector is examined. If the value of a feature is below a specific threshold, the the left branch is followed; otherwise, the right branch is follow. As the leaf node is reached, the decision is made about the class to which the example belongs.

- Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do and it generalises them by allowing optimisation of an arbitrary differentiable loss function. The objective of any supervised learning algorithm is to define a loss function and minimize it and this is the case for for Gradient Boosting algorithms.

- XGBoost is an implementation of Gradient Boosted Trees, popular in applied Machine Learning such as Kaggle competitions due to the algorithms execution speed and and model performance.

- XGBoost works by implementing the Gradient Boosting Tree algorithm:
  - Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made; and
  - Gradient Boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called Gradient Boosting because it uses a gradient descent algorithm to minimise the loss when adding new models.

- Consider the case where there are thousands of features, and therefore thousands of possible splits. Now, if we consider the potential loss for all possible splits to create a new branch we have thousands of potential splits and losses. XGBoost tackles the inefficiency of single Decision Trees by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits. Although XGBoost implements a few regularisation tricks, this speed up is by far the most useful feature of the library, allowing many hyperparameter settings to be investigated quickly. This is helpful because there are many hyperparameters to tune which are designed to limit overfitting.

- The reason for creating an additional, Gradient Boosted Decision Tree-based model using XGBoost was to build a Chance Quality model using raw shot location data (x and y coordinates), rather than engineered features such as angle and distance to the goal. This means that the Expected Goals predictions can be interpreted by referencing real positions on the pitch, rather than the more abstract distances and angles. This is not possible with logistic regression models without losing accuracy as logistic regression predictions are a linear combination of weights so factors with interactions, such as x and y coordinates, are more difficult to encode with linear weights.

- The most common argument against using Gradient Boosting algorithms such as XGBoost is, that it is a ”black box”. However, as with Decision Trees, the model’s features can be interpreted by packages such as the SHAP package to determine which are the most important features.

[1]XGBoost official documentation: https://xgboost.readthedocs.io/en/latest/
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

Model Selection 3/4 – Advantages of Logistic Regression

Strengths and weakness of using Logistic Regression for classification

Advantages

- Logistic Regression is very easy to implement.
- Logistic Regression is easier to interpret than XGBoost. The model coefficients of Logistic Regression can be used to indicate the importance of the features – no need for extra libraries such as SHAP.
- Fast training speed - faster than XGBoost.
- More stable model than XGBoost.
- Logistic Regression is already well calibrated i.e. when the probability of classification is 0.3, that represents a 30% chance of a goal being scored. Unlike Logistic Regression, XGBoost requires calibration after training.
- Logistic Regression is less inclined to overfit than Decision Trees and Gradient Boosting algorithms.
- A well-tuned Logistic Regression model can perform nearly as well as a model created using Decision Trees or Gradient Boosting algorithms such as XGBoost.

Limitations

- Logistic Regression regression doesn't deal well with non-linearities unless you add higher order terms. For this reason, domain knowledge and feature engineering is required to encode non-linear relationship. In the case of a Chance Quality Model, instead of using the x and y location on the pitch, the distance from the goal was determined and used as the feature to determine the shot-taker's location on the field.
- Less accurate than XGBoost. Many Kaggle competitions are won using XGBoost, not Logistic Regression. However, XGBoost is not guaranteed to be better than Logistic Regression in every setting.

Logistic Regression wiki: https://en.wikipedia.org/wiki/Logistic_regression.
Notebook to create the Chance Quality Model from the shots data using Logistic Regression (initial model):
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Model Selection 4/4 – Advantages of XGBoost

Strengths and weakness of using XGBoost for classification

Advantages

- XGBoost is more accurate than Logistic Regression. Many Kaggle competitions are won using XGBoost, not Logistic Regression. However, XGBoost is not guaranteed to be better than Logistic Regression in every setting.
- XGBoost handles no linear non-linearities better than Logistic Regression. In the case of a Chance Quality Model, XGBoost doesn't require that x and y location on the pitch are converted to a distance from the goal when determining the shot-takers position.
- XGBoost can handle categorical features without using one-hot encoding to encode features since several tests can combine to split a categorical feature.
- Implementation of XGBoost is also relatively easy and it's easy to get a good performance with little tuning.
- XGBoost is designed to handle missing data with its in-build features (sparse-aware). In the case of this Chance Quality Model, all missing values were treated and this was not a requirement from the chosen algorithm.
- Interpretation of the features is possible with external libraries such as SHAP.
- The user can run a cross-validation after each iteration.
- It supports regularisation.
- It works well in small to medium dataset.

Limitations

- XGBoost requires to be calibrated once the model is trained. i.e. in the case of a Chance Quality Model, when the probability of classification is 0.3, that represents a 30% chance of a goal being scored in a Logistic Regression model. However, in the case of XGBoost, without classification, this is not the case.
- Decision Trees generally don't extrapolate well to factors in the modelling that are hidden or not available in the dataset. Logistic Regression however is better at handling this.
- Decision Trees and Gradient Boosting algorithms are more inclined to overfit.
- Less easy to interpret than Logistic Regression (an extremely important aspect of football data analytics is interpreting the model results to football practitioners). While Decision Trees are also interpretable because of their strict if/else tests, the ensemble methods, which have higher generalisation ability, do not have a straightforward interpretation.
- Slower training speed than Logistic Regression Tuning hyperparameters using search functions such as Grid Search can take some time.

XGBoost official documentation: https://xgboost.readthedocs.io/en/latest/.
Notebook to create the secondary Chance Quality Model using XGBoost (separate notebook to Logistic Regression:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/2)%20Gradient%20Boosted%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Metric Selection 1/2 – All Available Metrics

There is an array of metrics available for classification model for which Log Loss is used in this model

- 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 e.g. position of the shot, left foot, number of defenders, etc. Therefore, the a loss function used needs to be appropriate for measuring the probability of a binary classifier.
- There is an array of model evaluation metrics at the disposal of a classification algorithm. Notable metrics include the: Accuracy, Precision, F1-score, ROC AUC, Brier Score, and the Log Loss. A brief description for each metric is described below but for more information and for all available metrics, see the scikit-learn documentation [link].
- In the following definitions, the following terminology is used:
  - True Positive (TP): Prediction is +ve and X is +ve
  - True Negative (TN): Prediction is -ve and X is -ve
  - False Positive (FP): Prediction is +ve and X is -ve
  - False Negative (FN): Prediction is -ve and X is +ve

Metric	Description
Accuracy	Accuracy is the ratio of the correctly labeled examples in the whole dataset of examples. Accuracy = (TP+TN)/(TP+FP+FN+TN).
Precision	Precision is the ratio of the correctly +ve labeled all +ve examples. Precision = TP/(TP+FP)
Recall (Sensitivity)	Recall is the ratio of the correctly +ve labeled out of all Recall = TP/(TP+FN).
F1-score	F1-score considers both Precision and Recall. It is the harmonic mean(average) of the precision and recall. F1-score = 2*(Recall * Precision) / (Recall + Precision).
Specificity	Specificity is the correctly -ve labeled that are actually -ve. Specificity = TN/(TN+FP)
ROC AUC	The ROC AUC of a model is the probability that the model ranks a random +ve example more highly than a random -ve negative example.
Brier Score	The Brier score, calculates the mean squared error between predicted probabilities and the expected values. The score summarises the magnitude of the error in the probability forecasts.
Log Loss (Binary Cross Entropy)	Log Loss or Binary Cross Entropy compares each of the predicted probabilities to actual class output which can be either 0 or 1. It then calculates the score that penalises the probabilities based on the distance from the expected value. That means how close or far from the actual value.

scikit-learn model evaluation metrics: https://scikit-learn.org/stable/modules/model_evaluation.html
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

Metric Selection 2/2 – Log Loss

Background information and mathematics to the Likelihood Function and the Log Loss

- This performance of the Chance Quality Model was assessed using the Log Loss, or the Binary Cross Entropy.
- The Log Loss is a take on the Likelihood Function (-1× the log of the Likelihood Function). The Likelihood Function answers the question of how likely did a model think the actually observed set of outcomes was. For example, if you have a labelled example in the training data and the model has found some values for the parameters, $\hat{\mathbf{w}}$, and the bias (real number), $\hat{b}$, you then apply the model $f_{\mathbf{w},b}$ to x_i and a value of $0 < p < 1$ will be returned. If y_i is the positive class, the likelihood of y_i being the positive class, according to the model, is given by p. Similarly, if y_i is the negative class, the likelihood of it being the negative class is given by $1 - p$.
- The optimisation criterion in Logistic Regression is called maximum likelihood. Instead of minimising the average loss, like in Linear Regression, Logistic Regression maximises the likelihood of the training data according to the model, represented mathematically as as:

$$L_{\mathbf{w},b} \overset{\text{def}}{=} \prod_{i=1\ldots N} f_{\mathbf{w},b}(\mathbf{x}_i)^{y_i}(1 - f_{\mathbf{w},b}(\mathbf{x}_i))^{(1-y_i)}$$

where:
- $\mathbf{w}$ is a D-dimensional vector of parameters
- b is bias (a real number)
- $f_{\mathbf{w},b}$ means that the model is parametrised by two values, $\mathbf{w}$ and b

- As Log Loss is -1× the log of the Likelihood Function, it can be formally defined as the negative average of the log of corrected predicted probabilities, and can be represented mathematically as follows:

$$\text{Log Loss} \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} -(y_i * \log(p_i) + (1 - y_i) \times \log(1 - p_i)$$

- Data Scientists used the Log Loss because for each prediction is between 0 and 1, ifyou multiply enough numbers in this range, the result gets so small that computers can't keep track of it. By taking the log of the Likelihood and multiplying it by negative 1, it is easy to keep track of and maintains a common convention that lower loss scores are better.

What is Log Loss by Dan Becker: https://www.kaggle.com/dansbecker/what-is-log-loss
scikit-learn model evaluation metrics: https://scikit-learn.org/stable/modules/model_evaluation.html
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

Treatment of Outliers

Full progression of handling unlikely and irregular shots in the dataset

A

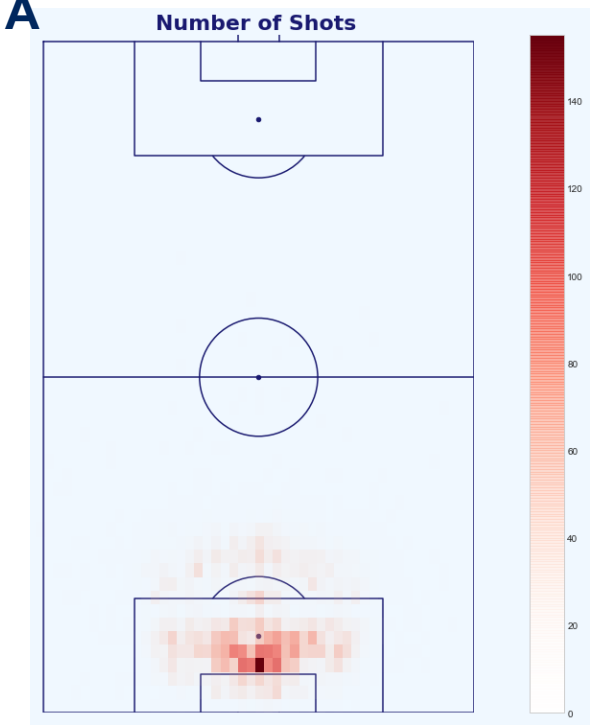


Fig. A: Heatmap of shots in the dataset, the center of the goal outside the 6-yard-box the most frequent.

B

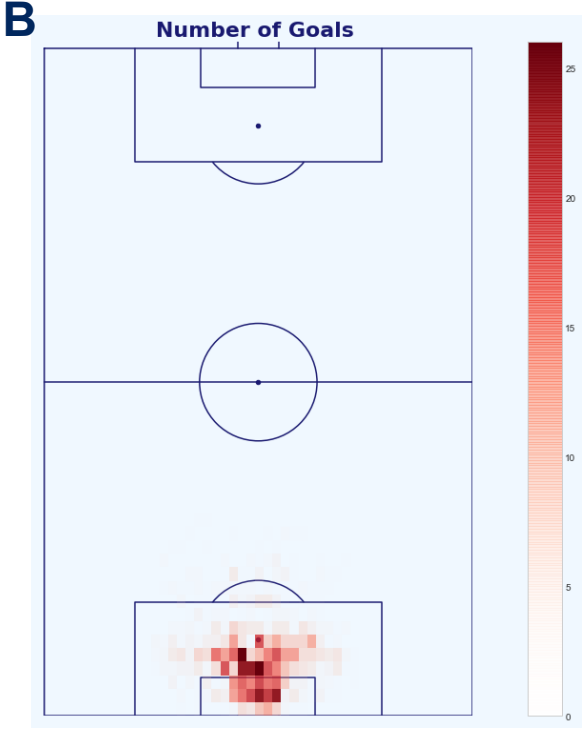


Fig. B: Heatmap of goals in the dataset, most scored within the 6-yard/18-yard box.

C

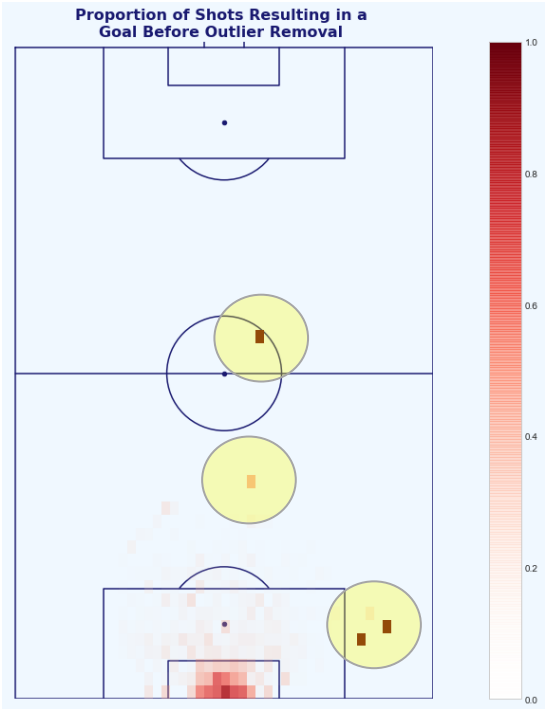


Fig. C: Heatmap of the proportions of shots to goals in the dataset (goals divided by shots). This visualisation flags some of the outlier shots scored from inside the attacking teams half, and at acute angles to the left of the keepers left post.

D

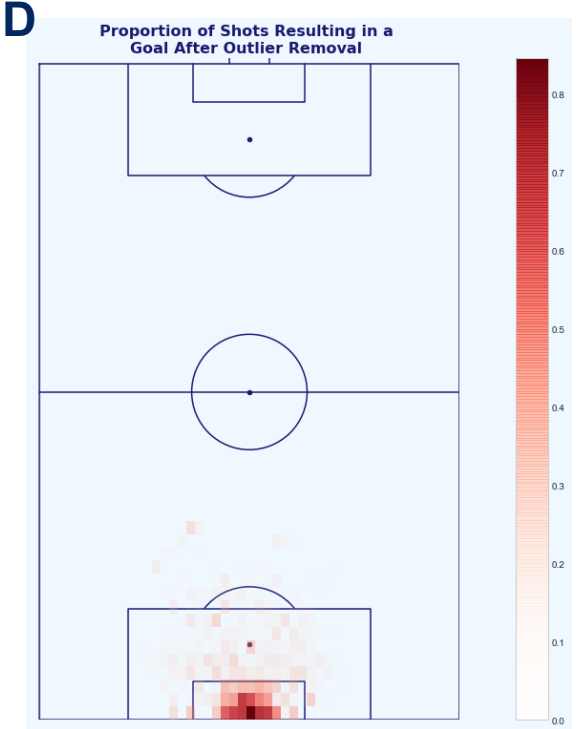


Fig. D: Heatmap of the proportions of shots to goals in the dataset post-outlier treatment. These outliers have now been removed to prevent this affecting the Chance Quality model.

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 Selection 1/3 – Selected Features for the Final Model

Description and status of features in the Chance Quality model from which predictions of the Metrica Shot data were made

Feature	Used in Final Model	Description
distance_to_goalM	Yes	Continuous feature that determines the distance the shot was taken along the y-axis in relation to the goal, in meters.
distance_to_centerM	No	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	Yes	Continuous feature that determines the angle in which the shot was taken to the goal.
number_intervening_opponents	Yes	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	Yes	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	Yes	Boolean feature that indicates whether the shot was taken with the player's foot, or not.
is_head	No	Boolean feature that indicates whether the shot was taken with the player's head, or not.
is_high_interference	Yes	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	No	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	Yes	Boolean feature that indicates whether the shooter experience a High level of interference (no or minimal interference).
header_distance_to_goalM	Yes	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

Features Selection 2/3 – Wishlist Features with Complete Event Data

A selected of key features I would like to include in an a CQM 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. This can be used as part of determining the assist type.
is_from_cross	Boolean feature to indicate whether a goal was scored from a cross. This can be used as part of determining the assist type.
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.
is_1_on_1	Boolean feature to indicate whether the shot was taken from a 1-on-1 situation i.e. the shot taker just has the goalkeeper to beat.
game_state	Categorical feature to indicate whether the shooting team is winning, drawing, or losing. As Michael Caley noted in his Expected Goals model, features for the game state (e.g. winning) appear be capturing latent factors that cannot be observed in the event data, such as the amount of defensive pressure asserted at the time the shot is taken.

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
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 Selection 3/3 – 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

XGBoost Modeling

Brief Summary of the steps taken in the XGBoost modeling

Hyperparameter Optimisation

- The optimum hyperparameters for the XGBoost model can be determined in several ways, two of the most common being Random Search and Grid Search. The parameters that this model focussed are were the ‘n_estimated’ (no. of Decision Trees used to train the model), the ‘max depth’ (maximum depth allowed for any individual tree), the ‘min_samples_split’ (the minimum number of samples allow to split a leaf into a new leaf), the ‘criterion’ (used to measure the quality of the split), and the ‘max_features’ (the maximum number of features allowed to use when determining how to split a node). For more information of XGBoost hyperparameters in scikit-learn, see [link].

Cross Validation

- Cross Validation is a method to ensure that the model finds the optimal set of parameters allowing for the model to fit the data as well of possible and generalise well to new data in the test set.
- A commonly used form of cross validation is k-fold cross-validation, used in this model, that splits the model k times with the relevant hyperparameters. For each run, one fold is set aside for a test set and k-1 folds are used for training the data, This prevents the overfitting of the test set by repeatedly training on the same data, The k results are averaged to produce a single estimation.

Feature Importance and Interpretation

- As previously mentioned, feature importance is extremely important in football analytics as just creating a model with a low Log Loss is no use when explaining the models to football practitioners. Clear, interpretable features are vital so that football can believe the results of the model.
- Feature importance in Gradient Boosted Decision Trees can be determined using the shap library or using Permutation Importance in the scikit-learn library. Using this, it was observed that the features that were most important to the model were the distance in which the shot was taken, the shot angle, and whether the shot was a header or not, as was observed in the Logistic Regression model.

This is just a brief summary of what was done in the secondary model. For more information about the Gradient Boosted Chance Quality Model and how the modeling took place, see the separate notebook [link].

XGBoost official documentation: https://xgboost.readthedocs.io/en/latest/.
SHAP library: https://github.com/slundberg/shap.
Permutation Importance in scikit-learn: https://scikit-learn.org/.
Notebook to create the secondary Chance Quality Model using XGBoost (separate notebook to Logistic Regression:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/2)%20Gradient%20Boosted%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb.

Model Performance

Iterations of the model during the modeling process

Model	Log Loss	ROC AUC	Accuracy*
Initial	0.33182	72.8%	88.5%[1]
Outlier Removal	0.32670	72.8%	-
Univariate Analysis	0.28979	80.7%	-
Multivariate Analysis	0.28892	80.7%	-
Final Model LR	0.28924	80.6%	-
Secondary Model (XGBoost)	0.28600[2]	-	-

[1]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
[2]After the initial submission, a secondary model has created, this time using XGBoost. See the following notebook:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/2)%20Gradient%20Boosted%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Exported Metrica Sports Shots 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	xG_LR	xG_XGB	diff_xG	percentage_variance_xG
2	57	Missed	39.22	18.36	22.96	53.11	18.36	0	1	0	Home	1	4,419	4,443	3	0	Medium	1	OFF TARGET-OUT	0	0	0.023	0.009	0.014	-60.49%
8	8	Goal	47.7	-3.4	6.30	32.68	3.4	1	1	0	Home	1	12,202	12,212	1	0	High	2	ON TARGET-GOAL	0	0	0.235	0.210	0.025	-10.63%
10	59	Saved	41.34	11.56	16.42	44.75	11.56	0	1	0	Home	1	16,484	16,499	2	0	Medium	1	ON TARGET-SAVED	0	0	0.091	0.068	0.024	-25.93%
12	21	Blocked	31.8	10.2	23.53	25.69	10.2	0	1	0	Away	1	18,515	18,520	1	1	Low	0	BLOCKED	0	0	0.083	0.081	0.002	-2.47%
18	14	Missed	43.46	-6.12	11.33	32.68	6.12	0	0	1	Home	1	27,345	27,360	1	1	High	2	BLOCKED	0	0	0.037	0.060	-0.022	59.73%
19	50	Saved	31.8	-13.6	25.19	32.68	13.6	0	1	0	Home	1	29,754	29,777	3	1	Low	0	HEAD-OFF TARGET-OUT	0	0	0.027	0.007	0.021	-76.18%
35	22	Goal	44.52	2.72	8.91	17.78	2.72	1	1	0	Away	1	53,049	53,075	1	0	Medium	1	ON TARGET-GOAL	0	0	0.379	0.282	0.097	-25.59%
37	23	Missed	42.4	6.12	12.24	30.00	6.12	0	0	1	Home	1	56,079	56,131	2	0	Low	0	HEAD-OFF TARGET-OUT	0	0	0.052	0.076	-0.024	46.89%
42	14	Missed	45.58	0.68	7.45	5.24	0.68	0	0	1	Away	1	63,362	63,372	1	0	Medium	1	OFF TARGET-HEAD-OUT	0	0	0.252	0.176	0.076	-30.25%
43	11	Saved	28.62	-3.4	24.62	7.94	3.4	0	1	0	Away	1	64,772	64,799	3	1	Medium	1	ON TARGET-SAVED	0	0	0.033	0.050	-0.017	50.62%
44	43	Missed	39.22	-18.36	22.96	53.11	18.36	0	1	0	Home	1	67,067	67,107	2	0	Medium	1	OFF TARGET-OUT	0	0	0.035	0.010	0.025	-70.76%
46	35	Missed	30.74	10.88	24.78	26.05	10.88	0	1	0	Away	2	69,887	69,927	2	1	Low	0	OFF TARGET-OUT	0	0	0.047	0.025	0.022	-47.65%
49	19	Goal	47.7	1.36	5.47	14.39	1.36	1	0	1	Home	2	73,983	73,995	2	0	Medium	1	HEAD-ON TARGET-GOAL	0	0	0.251	0.363	-0.111	44.18%
57	28	Missed	32.86	2.72	20.32	7.69	2.72	0	1	0	Away	2	86,191	86,219	4	2	Low	0	OFF TARGET-OUT	0	0	0.050	0.088	-0.038	75.32%
60	7	Saved	46.64	-11.56	13.19	61.18	11.56	0	1	0	Away	2	90,165	90,176	2	0	Low	0	ON TARGET-SAVED	0	0	0.135	0.079	0.055	-41.17%
65	55	Saved	48.76	-4.08	5.88	43.90	4.08	0	1	0	Home	2	98,880	98,896	1	0	Low	0	ON TARGET-SAVED	0	0	0.450	0.419	0.032	-7.01%
74	30	Blocked	43.46	-7.48	12.12	38.10	7.48	0	1	0	Away	2	111,758	111,763	3	1	Low	0	BLOCKED	0	0	0.143	0.154	-0.011	7.46%
76	40	Goal	40.28	1.36	12.79	6.10	1.36	1	1	0	Away	2	115,009	115,024	1	0	Low	0	ON TARGET-GOAL	1	0	0.760	0.760	0.000	0%
78	9	Missed	41.34	7.48	13.85	32.68	7.48	0	0	1	Home	2	117,218	117,245	2	0	Medium	1	HEAD-OFF TARGET-OUT	0	0	0.028	0.042	-0.014	49.35%
80	41	Goal	30.74	5.44	22.92	13.73	5.44	1	1	0	Home	2	121,027	121,055	2	0	Medium	1	ON TARGET-GOAL	0	0	0.054	0.074	-0.020	36.49%
82	53	Missed	45.58	0	7.42	0.00	0	0	0	1	Home	2	124,336	124,365	3	0	Medium	1	HEAD-OFF TARGET-OUT	0	0	0.133	0.259	-0.125	94%
88	23	Missed	33.92	-8.16	20.75	23.16	8.16	0	1	0	Away	2	132,570	132,597	6	4	Low	0	OFF TARGET-OUT	0	0	0.020	0.053	-0.032	158.68%
90	42	Missed	47.7	-2.72	5.96	27.17	2.72	0	0	1	Away	2	136,060	136,078	2	0	High	2	HEAD-OFF TARGET-OUT	0	0	0.095	0.071	0.025	-25.87%
93	16	Saved	21.2	1.36	31.83	2.45	1.36	0	1	0	Home	2	139,891	139,925	4	0	Low	0	ON TARGET-SAVED	0	1	0.009	0.008	0.001	-15.01%

Next Steps to Improve the Chance Quality Models

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, a simple Chance Quality Model was built using Logistic Regression. However, Gradient Boosting algorithms lead to improved performance and for this reason, a second model was made after the initial submission, creating a Chance Quality Model using XGBoost. Potential further models that can be deployed to try and further improve the performance of the Chance Quality model include other Gradient Boosting algorithms, such as LightGBM, and CatBoost. 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.

- The interference of the shot-taker in the provided dataset of shots was not defined in the documentation. For the Metrica Sports data, I have used the same definition that StatsBomb uses in their for their interpretation of a pressured player, in which an opposing player enters the 5-yard-radius of the player with the ball. However, I do not know how this stacks up to what a pressured player is in the Shots dataset and this will almost certainly be affecting the calibration of this feature in the dataset.

- 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.

- The model created only considers Open-Play shots. As per Michael Caley's model, it would be interesting to create Expected Goals models that also include Direct Free Kicks and Corners. In Caley's model, he creates six models for different match situations to reflect the varying difficult of shots in certain scenarios, these include: regular shots, shots from a direct free kick, headed shots from a cross, headed shots not from a cross, non-headed shots from a cross, shots following a dribble from the keeper thus the goalkeeper is not in goal when the shot is taken. Currently the xG value for penalties was taken from StatsBomb/FBref [link].

- Consider including more football knowledge in the model by adding fake shots to the dataset. As David Sumpter explains in the following tweet [link], adding fake data can improve the performance of the model and the limitations of Event data for two reasons – 1) it allows you to include things you know that are impossible (put players never do because its impossible) and 2) you can push the non-linear terms to really understand how the probability of success is shaped. For example, this could be done by adding in data which says it is impossible to score from the goal line or from more than 45m.

Michael Caley's xG model: https://cartilagefreecaptain.sbnation.com/2014/9/11/6131661/premier-league-projections-2014#methoderology
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

Advantages and Challenges of Event data

Pros and cons of the application of Event data for the analysis of football matches

Advantages

- Widely available for hundred of competitions from a variety of providers – StatsBomb, Wyscout, Opta.
- Increasingly information rich – new metrics included each season i.e. StatsBomb 'Freeze Frame' metric and StatsBomb's recently announced 360 update [link] that will collect a freeze-frame of all players on camera for every event recorded.
- Publicly available Event data has been available for a few years, leading to a number of libraries and repositories including socceraction (Python) by SciSports and ggsoccer (R) by Ben Torvaney.
- Easier to process and analyse than Tracking data. Only one event at a time and non-continuous. Size of the data is just 10^3 rows.
- Easy to connect Event data to video e.g. drag and drop an Opta F24 XML feed to SportsCode and query the video using Event data.

Challenges and Limitations

- Different providers have different interpretations of actions and accuracies of moments analysed. Subsequently, these datasets require different treatments. Libraries are available to transform data to a uniform standard such as SciSports SPADL format.
- Event stream is not always optimised for analysis.
- Event data is missing a lot of contextual information when compared with Tracking data as it only account for on-the-ball actions. Fewer than 1% of the on-the-ball actions in a football match are shots and players only have the ball 3 minutes on average (Johan Cruyff)[1]. Event data is unable to answer the questions as to what the players are doing during the 87 minutes when you do not have the ball.
- Not always pleasant to work with in its native form e.g. nested JSON dictionaries.

Cruyff quote taken from both Javier Fernández and Luke Bornn's paper 'Wide Open Spaces: A statistical technique for measuring space creation in professional soccer'.
Advantages and challenges of Tracking data discussed in Sam Gregory and Devin Pleuler's Sport Logiq webinair on 4th November 2019 'Demystifying Tracking Data'. See the following link for this seminar on YouTube: https://www.youtube.com/watch?v=miEWHSTYvX4.

Advantages and Challenges of Tracking data

Pros and cons of the application of Tracking data for the analysis of football matches

Advantages and Possibilities
(many still to be discovered!)

- A much fuller dataset, including all activity off-the-ball.
- Can take into account velocities, accelerations, and variables that provide continuous context, not just over a single frame e.g. a player's off-the-ball run; the ball trajectory on a shot etc.
- Data can provide tactical context from coordinate positions e.g. how many times is player X unmarked; where is the fullback when a transition occurs; provide another example of when event X occurred, etc.
- Tracking data can be used to further enrich corresponding Event data to add further detail and specificity. In the recent StatsBomb 360 Event data, Event data will now include the positions of all players for a given event. Other example of data enrichment include through computer vision and Tracking data include:
  - Is a pass under pressure? – distance based logic based on player in possession;
  - What passing options did the player in possession have – can be calculated geometrically or more simply through rules-based logic such as no. teammates and opponents in front or behind a player; Improve on existing metrics such as 'Packing[1]' which is currently manually collected;
  - Improved xG models to include defender and goalkeeper positions, pressure on the player in possession; and
  - Improve on previously hard-to-analyse metrics such as goalkeeper performance e.g. positioning at the moment of the shot, percent of the goal covered, etc.

Challenges

- Size of the data – Event data is 10^3 rows. Tracking data is 10^6 rows i.e. 95 mins (5,700 seconds) x 26 objects (22 players + 3 referees + ball) x 25 frames per second ≈ 3.7mil observations. This brings issues of both storing and processing the data.
- Mathematically more complex to work with.
- Despite being the state-of-the-art dataset, Tracking data is still missing significant datapoints including: body pose position (i.e. are they open to receive a pass), the spin of the ball (therefore need to build in uncertainties when determining ball trajectory).
- Limited public work - recently changed with the Friends of Tracking initiative + public datasets released by Metrica Sports, Signality, and SkillCorner. However, nearly all public initiatives are presented in journals or conferences and are therefore less accessible and aimed at a different audience to most football practitioners.
- Difficult to merge Tracking and Event data. Event data has imperfect timestamps. This is possible through a supervised modelling approached (requires properly annotated dataset) or a rules based logic approach (less accurate but easier to implement).
- Difficult to query Tracking data as the data is continuous and the unlike the Event data, the rows continue simultaneously with 26 things at the same time. For example, it's only possible to show all the passes if you've previous defined them.

[1]Packing' is a metric invented by IMPECT that assigns a value to the number of opposition players taken out of the game by a pass or dribble.
Advantages and challenges of Tracking data discussed in Sam Gregory and Devin Pleuler's Sport Logiq webinair on 4th November 2019 'Demystifying Tracking Data'. See the following link for this seminar on YouTube: https://www.youtube.com/watch?v=miEWHSTYvX4.