Department of Business and Management

Bachelor's Degree in Management and Computer Science

Chair of Databases & Big Data

Redefining Objective Scoring Chances: A Comprehensive, Unbiased Machine Learning Approach to Football Expected Goals

Andrea Marcoccia

Supervised by Prof. Blerina Sinaimeri

ID: 257681

Academic Year 2022/2023

Table of Contents

1. Introduction

1.1 Football Analytics: A Game-Changing Revolution

In the last decade, analytics have permeated every industry, redefining traditional practices and creating data-driven decision-making paradigms. Football (or soccer, as it's known in some parts of the world) is no exception. Once a game of intuition and gut-feeling, modern football has undergone a significant transformation, ushering in an era where analytics play a crucial role in understanding the complexities of the game.

Analytics in football serve multiple purposes, ranging from player performance analysis and tactical scrutiny to injury prevention and market valuation. With the advent of technologies such as video analysis software, GPS tracking, and machine learning algorithms, football clubs, analysts, and coaches now possess a wealth of data at their fingertips. This quantitative approach has not only led to more effective strategies on the field but has also influenced scouting, talent identification, and even game-day decisions.

1.2 Expected Goals (xG): The Cornerstone of Modern Football Analytics

One of the most groundbreaking metrics that have emerged in football analytics is the concept of "Expected Goals" or xG. Expected Goals is a statistical measure that quantifies the probability of a given shot resulting in a goal. It takes into account various factors such as shot angle, distance from the goal, type of assist, and even the part of the body used to take the shot. The xG metric assigns a value between 0 and 1 to each shot, indicating the likelihood that it will result in a goal. For example, a shot with an xG value of 0.3 has a 30% chance of being converted into a goal.

The xG metric has become indispensable in modern football analytics for several reasons:

- **Performance Evaluation**: xG offers a more nuanced way to assess both individual player performance and team tactics.
- **Strategic Planning**: Understanding xG can guide teams in optimizing their offensive and defensive strategies.
- **Fair Assessment**: By considering the quality of chances created or conceded, xG provides a more balanced view of a match, beyond the basic scoreline.
- **Predictive Power**: The metric is also used for predictive modeling, helping to forecast future performances and outcomes.

By transcending the limitations of traditional statistics like shots or possession percentages, Expected Goals has become a cornerstone metric that provides a deeper, more accurate insight into the game's intricacies.

More information on it can be finded in the article of Soccerment on xG and other advanced metrics[1].

1.3 The Limitations of Conventional xG Predictive Models

In the burgeoning field of football analytics, machine learning models designed to predict Expected Goals (xG) often incorporate a variety of features. These can include metrics related to the player taking the shot, the team they belong to, or even the opposing team. While this approach may yield a model with high predictive power, it introduces a significant problem—bias.

When the goal is to predict whether a shot will result in a score, using attributes of the player or team might seem advantageous. However, this approach becomes problematic when the objective shifts to using the xG metric as a fair evaluation of individual or team performance. In such cases, the xG value becomes inherently biased, influenced not just by the quality of the shot opportunity but also by who took it or which team they belong to. This not only muddies the analytical waters but also undermines the very purpose of using a metric like xG to assess performance impartially. It would be like judging a movie's quality based on the lead actor rather than the story itself.

1.4 The Goal: Creating an Unbiased xG Metric

The primary objective of this thesis is to address the limitations of conventional xG predictive models by developing machine learning models that rely exclusively on situational features. This means that the predictors chosen for the models are intended to describe the circumstances under which the shot was taken, rather than attributes of the player or team involved.

Examples of these situational features include:

- **Position of the Shot**: The coordinates on the pitch where the shot was taken from.
- **Body Part Used**: Whether the shot was taken with the foot, head, or any other body part.
- **Type of Preceding Event**: What happened right before the shot? Was it a corner kick, a free kick, or perhaps a dribble?

These features aim to create an xG metric that isolates the quality of the shot opportunity, thereby providing a fair and universal ground for performance evaluation. This objective is crucial for several reasons:

- **Fair Evaluation**: By solely focusing on situational factors, the metric provides a level playing field for evaluating individual players and teams.
- **Analytical Purity**: The metric remains true to the core purpose of xG, which is to quantify the quality of goal-scoring chances.
- **Universal Applicability**: The exclusion of player and team attributes ensures the metric's reliability across diverse settings for comparative analyses.

The ensuing chapters will delve into the methodology, data analytics, and machine learning models that underpin this pursuit.

2. The Data

2.1 Source

One of the significant challenges in the realm of football analytics is the scarcity of publicly available data. Most datasets in this field are proprietary assets of private companies, which limits the scope for independent research and analysis. In contrast, the dataset utilized for this thesis is a notable exception. Sourced from Wyscout, a leading company in the soccer industry, the data have been made publicly available by Pappalardo et al.[2] under the CC BY 4.0 License on figshare.com. This offers a unique advantage for comprehensive study and analysis in football analytics.

2.2 Collection Methodology

Wyscout's approach to data collection is both rigorous and meticulous, carried out by a team of expert video analysts using a proprietary software known as "the tagger." As described by Pappalardo et al.[1] the process can be broken down into three main steps:

- **Setting Formations**: At the beginning of each match, operators set the starting formations for the participating teams, identify the players' positions on the field, and record their jersey numbers.
- **Event Tagging**: Every ball touch in the match is tagged, creating a new event on a timeline. The operator then adds various attributes like the type and subtype of the event (e.g., pass, duel, shot), as well as its precise coordinates on the pitch.
- **Quality Control**: Once the tagging is complete, a two-step quality control process is initiated. First, an algorithm automatically cross-checks the tagged data to identify and

correct errors. This is followed by a manual review to further ensure the data's accuracy.

2.3 Dataset Overview

The original dataset is a multifaceted resource, providing varied types of data ranging from match outcomes and player features to events, referees, coaches, and much more furnished in the JSON format. For the scope of this thesis, the emphasis is placed on the dataset pertaining to events, complemented by a mapping dataset for tag identifiers and tag names.

The data covers the 2017/2018 season of the top-tier national soccer competitions in five leading European countries. Specifically, these are: Serie A in Italy, La Liga in Spain, Ligue 1 in France, Bundesliga in Germany and Premier League in England.

Furthermore, the dataset extends to encompass significant international tournaments, specifically the World Cup 2018 and the European Cup 2016, which are competitions for national teams.

Our events dataset is not just rich in detail; it's also expansive in scale. It contains 3,251,294 records, and, according to Pappalardo et al.[2] in 2019, represents the largest public collection of soccer-logs ever released to the best of their knowledge. This sheer volume provides us with an incomparable opportunity to draw statistically significant conclusions and develop robust machine learning models for predicting Expected Goals (xG).

Each record provides a snapshot of a specific event in a football match, including the following features:

- **eventId**: An identifier for the type of event
- **eventName**: The name of the event's type (e.g. Pass, Shot, Duel, Foul)
- **subEventId**: An identifier for the sub-type of the event
- **subEventName**: The name of the sub-type of the event (e.g. Simple Pass, Cross, High Pass, Head Pass)
- **tags**: A list of tags providing additional information about the event. (e.g. Accurate, Not Accurate, Key Pass, Assist)
- **eventSec**: The time in seconds when the event occurs relative to the current half of the match
- **id**: A unique identifier for the event
- **matchId**: The identifier of the match to which the event belongs
- **matchPeriod**: The period of the match when the event occurs (e.g., 1st Half, 2nd Half)
- **playerId**: The identifier of the player who generated the event
- **positions**: The x, y coordinates indicating the event's position on the field
- **teamId**: The identifier of the player's team

Our primary interest in this dataset revolves around shots. This emphasis is because shots, more than any other event type, will serve as the cornerstone for our predictive models.

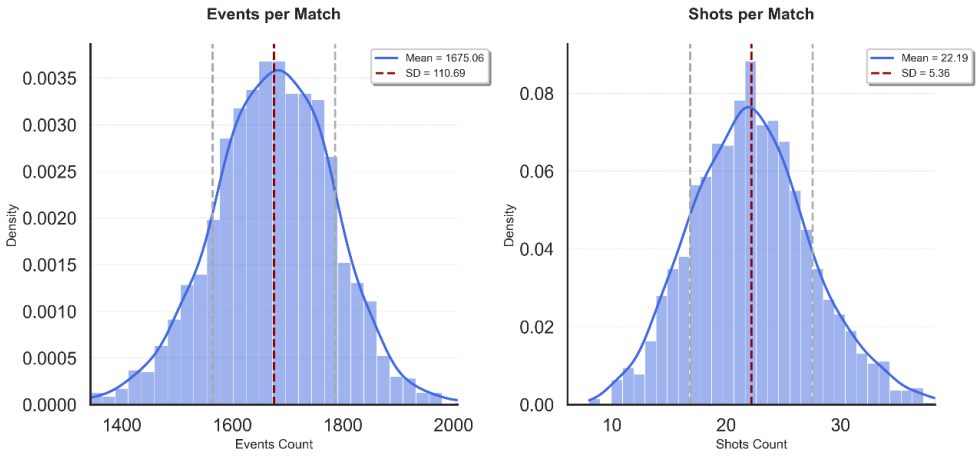

Figure 1: Frequency of Events and Shots

When we dive into the visual representation of our dataset in Fig.1, two patterns emerge. Both the frequency distributions for events per match and shots (from open play) per match mirror the characteristics of a normal distribution. The average match showcases approximately 1675 events, with a standard deviation of 110. Matches typically see an average of 22 shots and deviating by a standard deviation of 5.36. This suggests that most football games follow a common pattern with only a few standing out as exceptions.

2.4 Feature Engineering

To construct a model that accurately measures objective scoring chances without bias, it's essential to have the right features. While the raw Wyscout dataset provided a comprehensive view of football matches, specific modifications were necessary to tailor the dataset for our goal.

The first step was to concentrate on the events central to our model – the shots. By filtering the data, we retained only the events specifically labeled as shots. Additionally, to ensure a comprehensive capture of all potential goal-scoring opportunities, shots originating from free-kicks and penalties were also included. After this refinement, our dataset was narrowed down to 45,945 records.

The Wyscout dataset is enriched with tags that contain valuable information about each event. Accompanying the data, Pappalardo et al.[2] provided a CSV file that establishes a mapping between tag IDs and their corresponding names. By leveraging this mapping dataset, we decoded the tags associated with each shot, transforming them from mere IDs to more descriptive and informative labels. This process allowed us to unearth several crucial

attributes. For instance, we could ascertain if a shot stemmed from a counter-attack or if it was identified as a clear-cut opportunity. However, it's paramount to note our selective approach: tags that were directly related to the outcome of the shots, such as interceptions or blockages, were intentionally excluded to ensure that biases were not inadvertently introduced into our ensuing analysis.

By analyzing events grouped by player IDs, we calculated the occurrences of events per foot for each player. This allowed us to assign a 'strong foot' for every player. Using this data, a new binary feature was crafted to indicate if a shot was taken with the player's strong foot.

Interpreting the positional coordinates requires careful consideration of the context provided by the Wyscout API documentation[3].

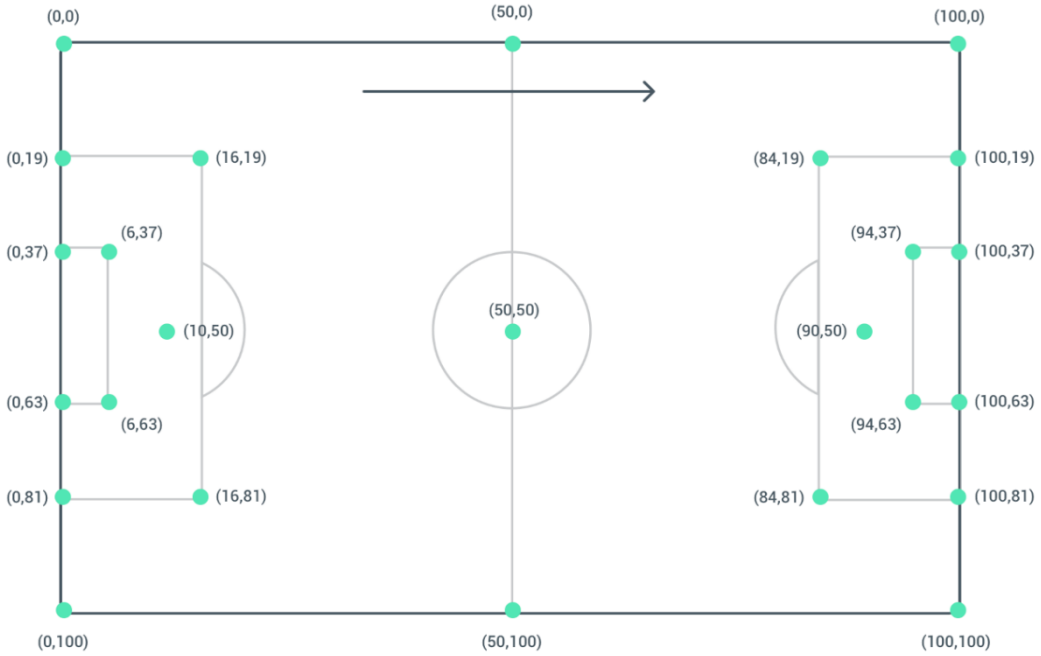

Figure 2: Visual representation of Coordinates

According to their guidelines, the coordinates are expressed as percentages, where the goal being defended is at x = 0% and the attacking goal is at x = 100%. Specifically, the center of the attacking goal is positioned at 100:50. Fig. 2 shows a visual representation of this.

Given this configuration, from the raw coordinates provided we computed two essential features: distance and angle of each shot from the center of the attacking goal.

In the pursuit of building robust predictive models for football analytics, it becomes imperative to blend depth of information with computational pragmatism. Delving deep into every event leading up to a shot undoubtedly furnishes richer context. However, it also brings forth challenges tied to complexity and computational demands. Recognizing these trade-offs,

we strategically anchored our focus on the immediate event preceding each shot. This nuanced approach not only ensured we garnered significant context but also maintained an efficient analysis.

Given the multifaceted nature of football matches, the events culminating in a shot can provide insights on the shot's prospective outcome. Thus, to gain insights into the circumstances immediately framing each shot, we turned our attention to the event preceding it. Employing a method where we grouped the data by matches and tapped into the time variable, we successfully pinpointed this precursor event. This exercise empowered us to extract and integrate pivotal features like:

- Type and subtype of the previous event
- Distance and angle of the previous event relative to the shot's position
- Time difference between the shot and its preceding event

To add depth to our understanding, we also looked at the tags associated with these preceding events. Given the myriad of possible events before a shot, there was a vast array of associated tags. While every tag provides a shade of context, not all would be vital for our predictive ambitions. To optimize our dataset without sacrificing critical nuances, we adopted a data-driven criterion: only tags that featured in more than 10% of the records were included. This strategy ensured that our model had access to prevalent and significant patterns, filtering out potential noise.

The final dataset for our analysis encompasses a diverse set of features, each offering unique insights into the dynamics of a football match:

Spatial Features:

- Distance: Continuous (0-100 meters). Measure the shot's proximity to the goal.
- Angle: Continuous (0-90 degrees). Indicates the shot's angle relative to the goal's center.

Temporal Features:

- Time: Continuous. Denotes seconds elapsed since the start of the current match period.
- Match Period: Categorical. Identifies the segment of the match (e.g., 1st Half, 2nd Half).

Shot Context:

- Situation: Categorical. Describes the context of the shot (Free Kick, Open Play, Penalty).
- Body Part Used: Categorical. Classifies the shot's execution method (Strong Foot, Weak Foot, Head/Body).
- Foot: Binary. Indicates which foot was used for the shot (Right or Left).

Wyscout Video Analysis Tags:

- Opportunity: Binary. Indicates if the shot was deemed an opportunity.
- Counter Attack: Binary. Identifies shots originating from a counter-attack.

Details of the Preceding Event:

- Event Type and Subtype: Categorical variables. Depict the nature of the event.
- Distance: Continuous (0-121 meters). Distance of the previous event from the shot.
- Angle: Continuous (0-90 degrees). Angle of the previous event concerning the shot.
- Inter-event Time: Continuous. Indicates the time gap between the shot and the previous event.
- Accurate: Binary. A tag marking the event as precise.
- Won: Binary. Present in duels, signifying if the duel was victorious.
- Key Pass: Binary. Present in passes, indicating if the pass led to a dangerous move.

Having refined and enriched our dataset, we've laid a robust foundation for the subsequent phases of this research. The careful selection and engineering of features are pivotal in ensuring the accuracy and reliability of our forthcoming predictive models.

3. Exploratory Data Analysis (EDA)

The process of data collection and preprocessing, though meticulous, is only half the battle. Equally, if not more important, is understanding the intricacies and patterns within the data. This chapter, the Exploratory Data Analysis (EDA), is where we embark on a journey to uncover those hidden insights, relationships, and anomalies. EDA serves as the bridge between raw data and the ensuing predictive models, ensuring we proceed with a well-informed perspective. Through visualizations, distributions, and statistics, we'll dissect our dataset, focusing on each variable's influence on a shot's likelihood to result in a goal.

3.1 Data Cleaning

The EDA commences by examining three pivotal variables: 'Situation', 'Match Period', and 'Distance'. These were selected as the initial focus because exploration of these variables influenced **data cleaning** decisions, particularly the removal of specific records. By starting with these features, the subsequent analyses are built on a more refined and relevant dataset.

Situation:

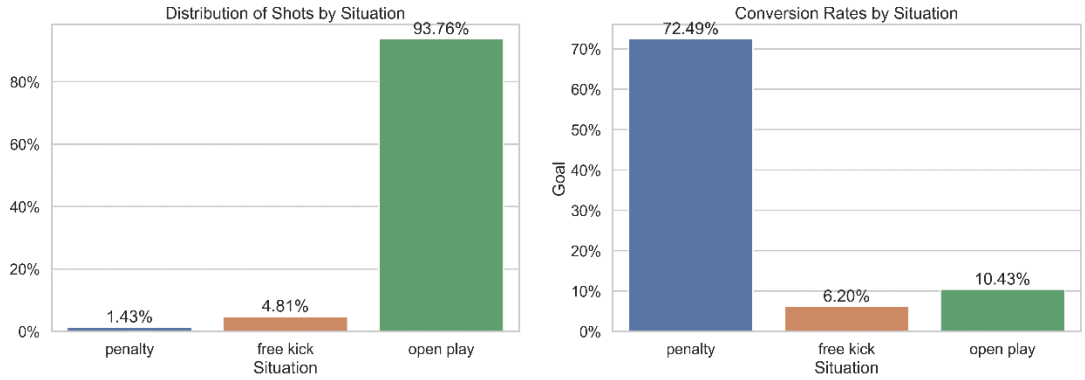

Figure 3: Distribution and Conversion Rate of Shots by Situation

Our exploration of the situation variable in Fig. 3 uncovers several key insights:

Distribution of Shots by Situation:

- A vast majority of the shots in the dataset come from open play, constituting 93.76%.
- Shots from free kicks represents 4.81%, while penalties comprise a smaller fraction, about 1.43%.

Conversion Rates by Situation:

- Penalties exhibit the highest conversion rate - 72.49%.
- Open play shots, despite their dominance in frequency, have a conversion rate of 10.43%.
- Free kicks, even with their strategic significance, yield the lowest conversion rate of 6.20%.

Given these insights, a strategic decision has been made regarding our modeling approach. Recognizing the fixed nature of penalties, it's prudent to assign a static Expected Goals (xG) value of 0.72 to all penalty shots. This decision is grounded in the consistent conversion rate observed for penalties. Consequently, our predictive models will be tailored exclusively for non-penalty shots. This ensures a more nuanced and adaptable xG model that can account for the varied contexts of open play and free kicks.

Match Period:

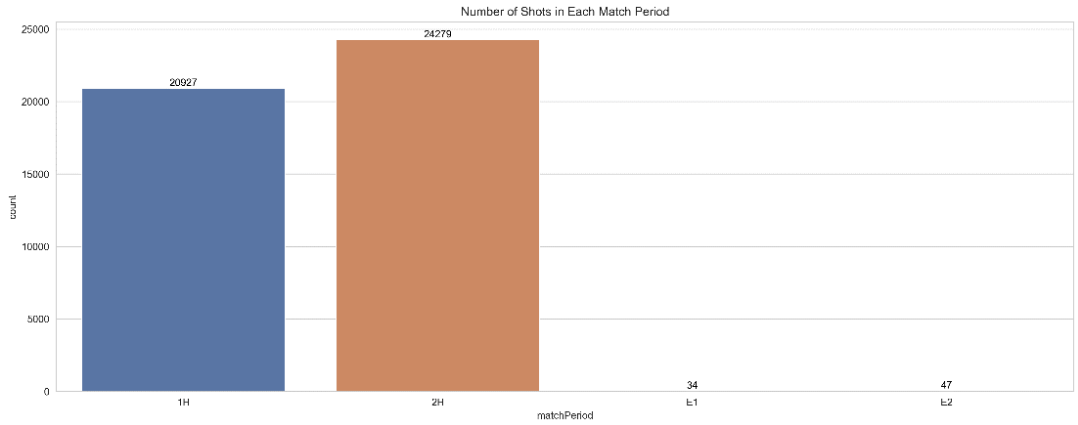

Figure 4: Distribution of Shots across different Match Periods

The distribution of shots across match periods shown in Fig.4 reveals 2 key insights:

- Dominance of Regular Time: As the plot illustrates, most shots are taken during regular time, specifically during the 1st and 2nd halves. Extra time (E1 and E2) contributes only a minimal amount to the total shots. This is consistent with the nature of football matches, given that extra time is conditional and occurs less frequently.
- Second Half Surge: The frequency of shots is notably higher in the second half compared to the first. This increase can be attributed to several factors. Tactical shifts often take place during half-time, potentially leading to more aggressive or open play. Additionally, player fatigue becomes more prominent as the match progresses, which can result in defensive vulnerabilities and, consequently, more shots. Lastly, the urgency of chasing a result, especially if a team is trailing, can result in a more forward-leaning approach.

Given the limited data from extra time periods (E1 and E2) it's prudent to focus our analysis on the regular time shots. This ensures a more robust dataset and avoids skewing results based on rare or exceptional circumstances. Hence, moving forward, our analysis and predictions will be based solely on shots taken during the 1st and 2nd halves.

Distance:

The distance from which a shot is taken plays a crucial role in determining its likelihood of resulting in a goal. To ensure the precision and relevance of our analysis, it's essential to address potential outliers in this variable. A boxplot was plotted to visually assess the distribution and identify outliers.

Figure 5: Boxplot of Shots across Distance to Goal

As seen in the boxplot in Fig. 5, the data exhibits a few extreme values that deviate significantly from the central tendency. To systematically address these outliers, the Interquartile Range (IQR) method was employed. This statistical technique defines an upper threshold, calculated as Q3+1.5×IQR. Any data points that lie beyond this threshold are considered outliers.

Based on this criterion, 241 records were identified as outliers and subsequently removed. While the decision to exclude data is always a significant one, it's important to note that these records represent a mere 0.5% of the entire dataset. Given this small proportion and the potential of these outliers to skew our analysis, their removal was deemed justified.

The analysis of the feature proceeds with the Kernel Density Estimate (KDE) of Goals and Not Goals across values of Distance.

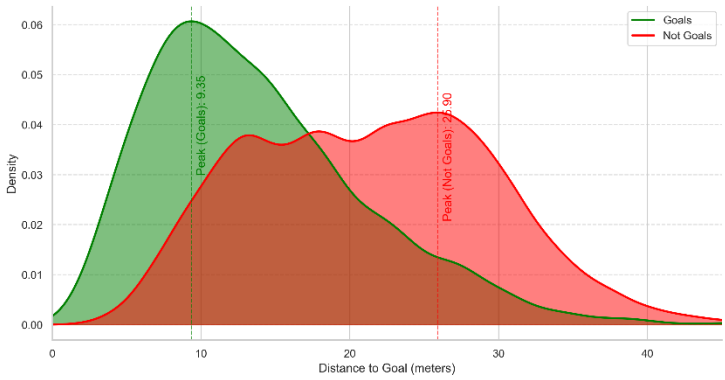

Figure 6: KDE of Goals and Not Goals across Distance

Two distinct distributions emerge from the kernel density plot in Fig. 6: one for shots that resulted in goals (green) and the other for shots that didn't (red). Several observations arise from this visual representation:

- Peak Distribution for Goals: The green curve, reaches its maximum at a distance of 9.35 meters from the goal. This pronounced peak highlights the most frequent distance from

which goals are scored, suggesting a "sweet spot" for players around this peak. Possibly due to being close enough to bypass defensive obstructions effectively.

- Peak Distribution for Not Goals: On the other hand, the red curve, peaks at about 25.90 meters. This distance represents the most common point from which players attempt to shoot, but fail to find the back of the net. It might hint at a zone where, despite the frequent attempts, defensive strategies, goalkeeper interventions, or the sheer difficulty of shooting from such a range diminishes the success rate.
- Shots decline after the peaks: For the goals, there's a significant decline in density post the peak, emphasizing the rarity and challenge associated with long-range goals. Similarly, the distribution for shots that didn't result in goals shows also decreasing density beyond the peak. This pattern indicates that while unsuccessful shots are more commonly attempted from longer distances than successful ones, teams regognize the decreasing odds of such shots leading to goals and make fewer attempts from these farther ranges.

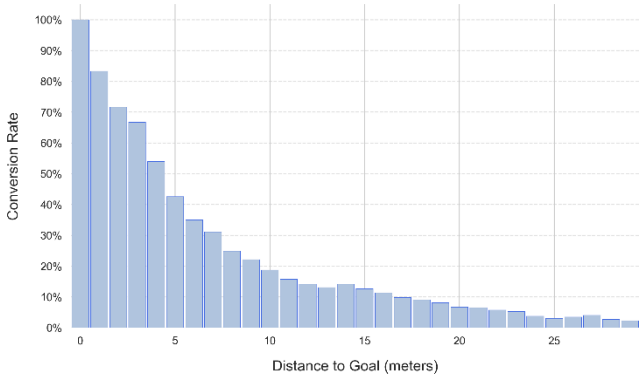

Figure 7: Conversion Rate across Distance

The histogram of conversion rates by distance in Fig. 7 supplements our earlier observations from the kernel density plot:

- Declining Conversion Rates: As we move farther from the goal, there's a notable decrease in the conversion rate. This trend confirms the challenges associated with long-distance shots.
- Proximity and High Conversion: For shots within 4 meters of the goal, the conversion rate surpasses 50%. This rate remains significant, over 20%, for shots up to 9 meters away.
- Approaching the Average: Around the 15-meter mark, the conversion rate aligns with the average rate of 10%. Beyond this point, the success likelihood further diminishes.

The clear patterns and relationships observed between distance and shot outcomes strongly suggest that this feature will play a pivotal role in shaping our Expected Goals (xG) predictions in subsequent models.

With the refined dataset in hand, the EDA progresses to the other features, beginning with an exploration of the remaining Categorical Variables.

3.2 EDA on Categoricals

Body Part:

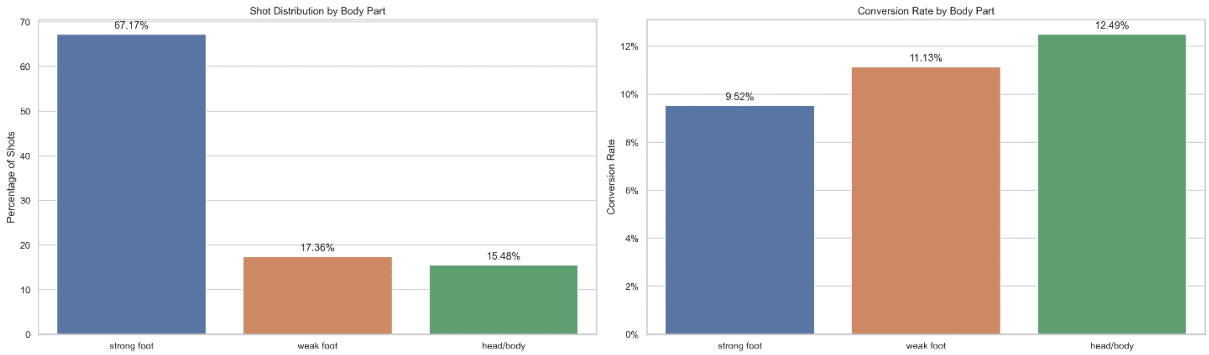

Figure 8: Shot Distribution and Conversion Rate by Body Part

The distribution of shots by body part in Fig. 8 reveals that the strong foot significantly dominates the scene, accounting for 67% of all attempts. Following this, the weak foot contributes to 17% of the shots, while headers or other body parts make up the remainder, rounding to 15%.

Diving into the conversion rates plot in Fig. 8, shots with the strong foot display a conversion rate of approximately 9.5%. Surprisingly, headers or shots taken with other parts of the body exhibit a slightly elevated conversion rate of around 12.5%. This could be a consequence of headers being more unpredictable for goalkeepers or originating from strategically advantageous set pieces.

Notably, despite the common expectation, shots taken with the weak foot yield a higher conversion rate of 11%, even surpassing those executed with the strong foot. This counterintuitive trend might be rationalized by understanding player behavior. Players might be inclined to shoot with their strong foot even under unfavorable conditions. In contrast, opting for the weak foot might be a choice made only when they find themselves in more favorable or unguarded positions, leading to higher success rates.

Foot:

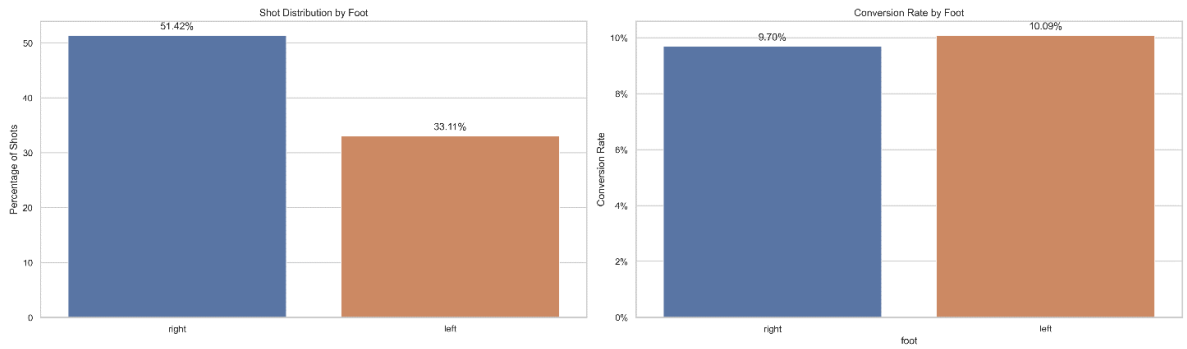

Figure 9: Shot Distribution and Conversion Rate by Foot

The graph in Fig. 9 distinctly shows a higher frequency of shots taken using the right foot as compared to the left. This trend is in line with the general observation in football, where players predominantly have a right strong foot. Specifically, approximately 51% of shots are taken with the right foot, while 33% are executed with the left foot.

Fig. 9 shows also comparable conversion rates for both feet, indicating that the foot's orientation isn't necessarily a decisive factor in converting a shot into a goal. The left foot, intriguingly, registers a slightly higher conversion rate, about 10%, compared to the right foot's 9.7%.

A potential rationale behind this observation is consistent with our earlier findings about body parts. As the left foot is more commonly the weak foot among footballers, the data might suggest that players tend to attempt shots with their weak foot under more favorable conditions. This, in turn, leads to a marginally better conversion rate for shots taken with the left foot.

Incorporating these findings, one can conclude that while the right foot is more commonly used for shooting, the success rate isn't heavily influenced by the foot's orientation. The subtle advantage for the left foot in terms of conversion rates further emphasizes the importance of situational context and decision-making in football.

Opportunity:

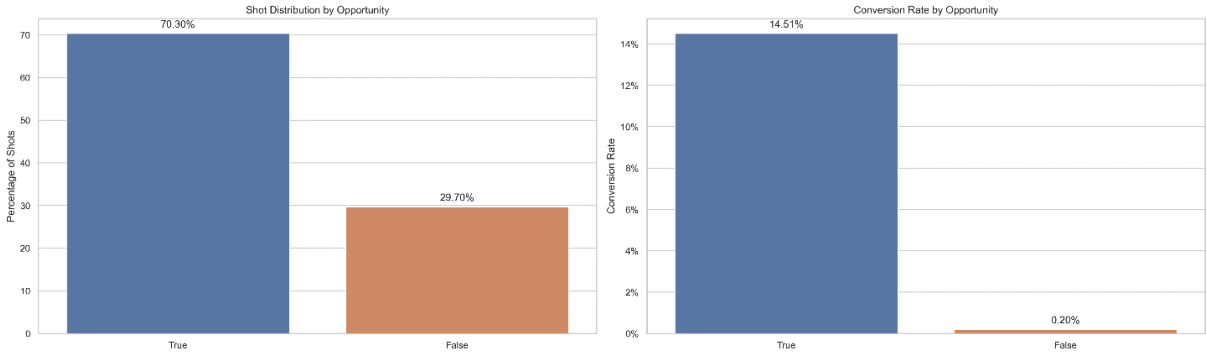

Figure 10: Shot Distribution and Conversion Rate by Opportunity

As shown in Fig. 10 the data delineates a clear disparity in shot frequency based on their categorization as an 'Opportunity'. Specifically, a substantial majority, approximately 70%, of shots are tagged as opportunities. In contrast, the remaining 30% don't bear this classification. This signifies that a significant chunk of the shots taken during matches arise from situations deemed favorable.

The conversion rates in Fig. 10 further accentuate the weight of the 'Opportunity' label. Shots that are labeled as opportunities have an impressive conversion rate of 14.5%. This rate is starkly higher than its counterpart, where shots not labeled as opportunities register a minimal conversion rate of just 0.20%. This vast chasm underscores the effectiveness and accuracy of the 'Opportunity' classification in determining the likelihood of a shot resulting in a goal.

Given this pronounced difference in conversion rates it's evident that this variable will play a pivotal role in the predictive modeling.

Counter Attack:

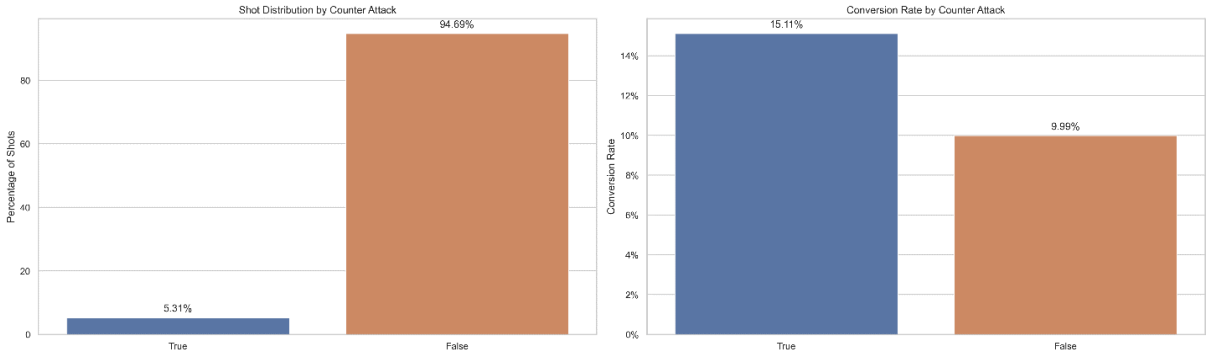

Figure 11: Shot Distribution and Conversion Rate by Counter Attack

Fig. 11 shows that counter-attack shots are relatively rare, representing just 5% of all shots. The remaining 95% of shots occur in other gameplay scenarios. This highlight the tactical nature of football, where organized attacks, rather than quick counters, dominate the play.

Fig. 11 shows also how counter-attack shots convert at a higher rate, approximately 15%, compared to the 10% conversion rate for non-counter attack shots. This suggests that counter attacks often lead to clearer goal-scoring opportunities, possibly due to catching the defense off-guard.

The data indicates that while counter attacks are infrequent, they are more potent in terms of goal-scoring potential.

Having delved into the categorical variables, the exploration now shifts to the continuous features of the dataset.

3.3 EDA on Continuous Variables

We commence this phase by examining the 'Angle' feature, which represents the angle at which the shot is taken relative to the goal.

Angle:

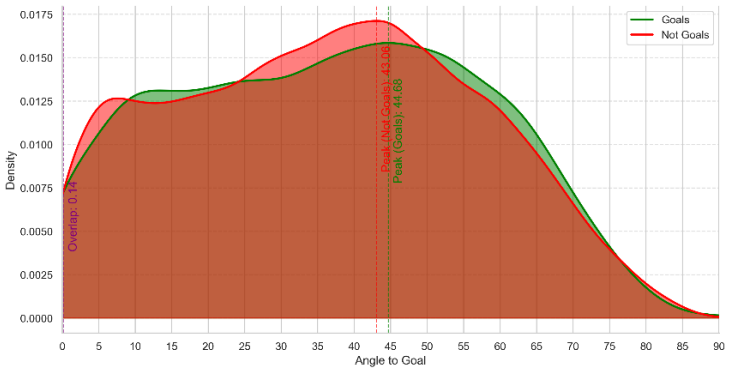

Figure 12: KDE of Goals and Not Goals across Angle

Two distinct distributions emerge from the kernel density plot in Fig. 12: one for shots that resulted in goals (green) and the other for shots that didn't (red). Several observations arise from this visual representation:

The peak of the distributions for both goals and non-goals are quite close to each other, with goals peaking at 44.68 degrees and non-goals at 43.06 degrees. Shots taken around these angles are common, regardless of the outcome (goal or not). The closeness of the peaks indicates that the probability of scoring or not scoring from these angles is quite similar.

Both distributions have very similar densities. The density stays up for angles up to around 60 degrees indicating that these angles are common for taking shots. Going over this threshold the density gets lower and lower, this suggests that players might find more challenging to score or even attempt shots from such extreme angles.

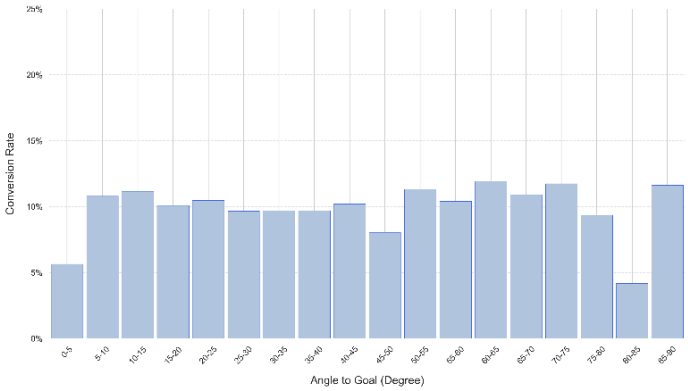

Figure 13: Conversion Rate by Angle

Fig. 13 shows that, across a wide range of angles, the conversion rate remains relatively consistent, fluctuating around the 10% mark.

However, there are 2 notable exceptions on the intervals in the plot:

- 80-85 degrees: The conversion rate drops significantly within this bin. However, it's crucial to interpret this with caution. Our earlier analysis demonstrated that the sample size for shots taken from such extremly big angles is quite limited. Therefore, the diminished conversion rate in this range might be more a reflection of the rarity of such events rather than a true indication of their improbability. The few cases present could disproportionately influence the observed rate.
- 0-5 degrees: Another interesting deviation is observed in the 0-5 degree bin, where the conversion rate is notably lower, hovering around 5%. The density for this angle bin, although low, remains significant. This suggests that shots from such narrow angles might inherently be more challenging, possibly due to factors like limited goal view or intensified defensive pressures.

From these observations, it's evident that the angle has a minimal impact on the likelihood of a shot resulting in a goal.

Time:

To gain a deeper understanding of the dynamics of the game with respect to shots taken, we've analyzed the 'Time' variable, representing the seconds elapsed since the start of each match period. For clarity, this data is presented in intervals of 5 minutes and is grouped by match period.

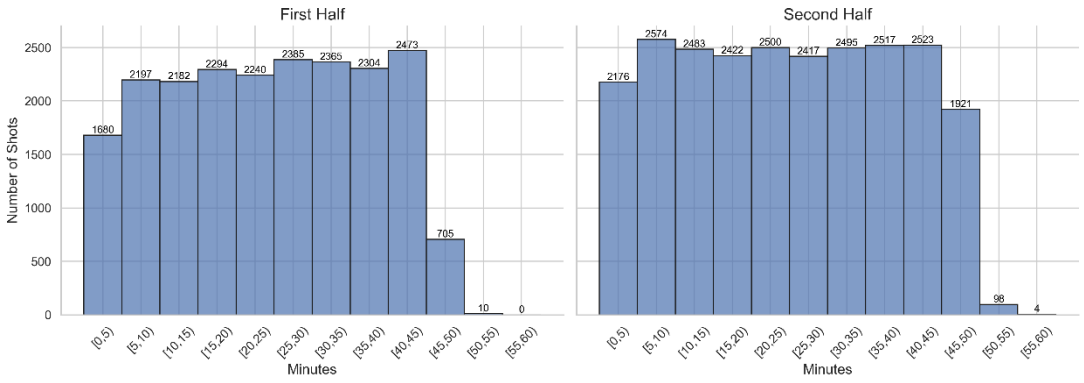

Figure 14: Shot Distribution over Time in 5-minute intervals by Match Period

The plot in Fig. 14 represents the frequencies of shots in our data at different intervals of the match and reveals different insights.

The first half of the game shows an increasing trend in the number of shots taken. This trend persists until the end of regular time at 45 minutes. Post the 45-minute mark, there is a noticeable dip in the number of shots. This reduction can be attributed to the variability of extra time, which can sometimes be just a few minutes, thereby limiting opportunities for teams to shoot.

The first 5 minutes appear to be a phase where teams are recalibrating, leading to fewer shot opportunities in both halves.

In the second half the number of shots doesn't display a consistent trend. Instead, it fluctuates around an average of approximately 2500 shots until the conclusion of regular time. Unlike the first half, the initial segment of the second half's extra time maintains a higher number of shots. This suggests that the second half generally has a longer extra time duration, allowing for more shot opportunities.

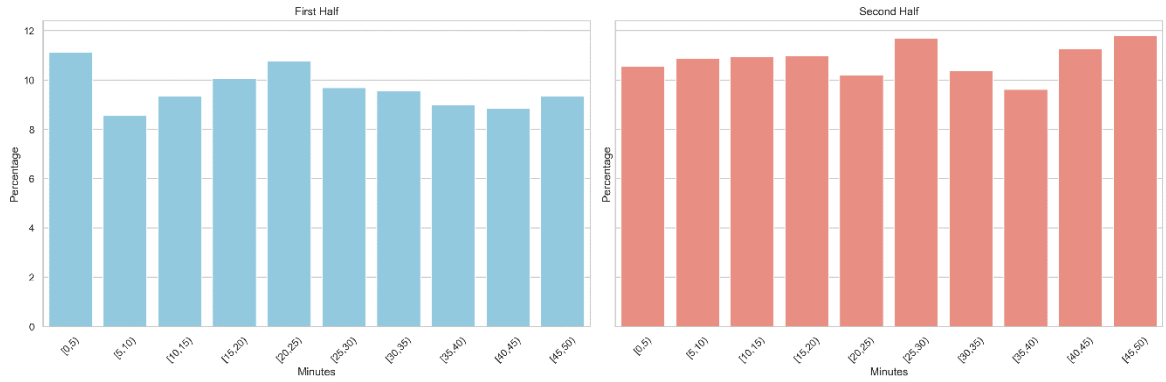

Figure 15: Conversion Rate over Time in 5-minute intervals by Match Period

The analysis of the Time variable poceed in Fig. 15 with a histogram of conversion rates at the different intervals up to the 50th minute. The decision of this threshold was driven by the low number of shots observed after this minute. Relying on a limited number of observations could introduce bias, making the derived insights potentially less statistically significant.

The conversion rate in the first half displays two noticeable peaks. An elevated conversion rate, surpassing 10%, is witnessed during the initial 5 minutes of the game. Another peak in conversion rate occurs midway through the first half, specifically in the [20-25) interval. For the remaining duration of the first half, the conversion rate predominantly remains below the 10% threshold.

Generally, the second half exhibits a superior conversion rate compared to the first half. The conversion rate consistently remains above 10%, with an exception noted in the [35,40) interval. The highest conversion rates in the second half are observed in the [25,30) interval and towards the end of the match, particularly in the intervals [40,45) and [45,50).

This analysis offers a comprehensive view of how the likelihood of converting a shot into a goal varies over different intervals during a match. The fluctuation in conversion rates at different times could be influenced by multiple factors such as team strategies, player fatigue, or the urgency to score as the match nears its conclusion.

3.4 EDA on Preceding Event

Upon detailed analysis, it becomes evident that the immediate dynamics of a shot—its angle, distance, etc—have a more pronounced influence on the outcome than the preceding event. This underscores the intricate and multifaceted nature of football, a sport where the immediate context of a shot often dictates its success more than the buildup play.

For the sake of clarity and to sustain reader engagement, the findings have been streamlined. Instead of delving into every nuance, the focus is shifted to highlight only the most salient and intriguing aspects discovered during the analysis. This approach ensures a concise yet informative exploration of the topic.

Impact of Previous Event Type vs. Event Sub-Type:

The relationship between the previous event type and the previous event sub-type is hierarchical. The sub-type is essentially a finer categorization within the broader event type. Because the sub-type inherently carries information about its parent event type, including both in a predictive model could introduce unnecessary redundancy. Thus, only one will be selected.

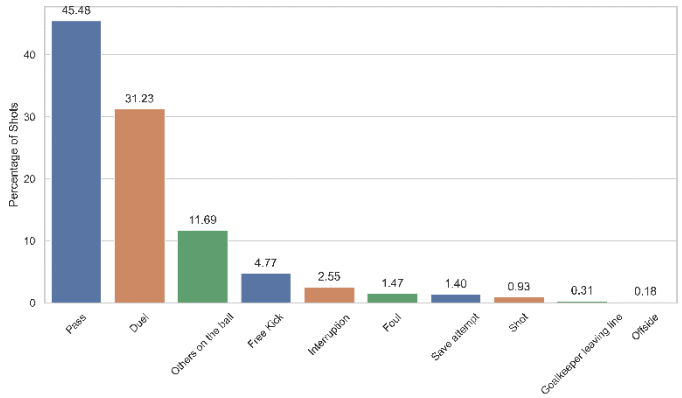

Figure 16: Shot Distribution by Previous Event Type

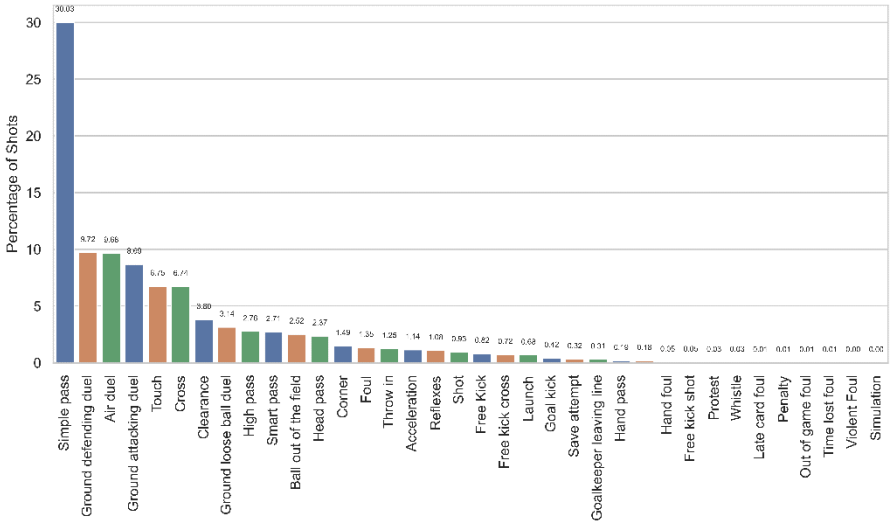

Figure 17: Shot Distribution by Previous Event Subtype

An examination of the frequency distribution of various categories, as depicted in Fig. 16 for the previous event type and in Fig. 17 for the previous event sub-type, provides valuable insights into the granularity of the data and its potential impact on model performance.

Observing Fig. 16, it's clear that while there is some variability in the frequency of different event types, categories maintain a significant presence in the dataset. Only three categories have frequencies below 1%, with the least frequent still accounting for 0.18% of the data. Given that the dataset encompasses over 40,000 records, even these less frequent categories are represented by a substantial number of instances.

Contrastingly, the frequency distribution of event sub-types in Fig. 17 paints a different picture. More than half of the categories are represented in less than 1% of the shots. Alarmingly, several categories dip to extremely low frequencies, with the rarest ones appearing in less than 0.01% of the records. This implies that these categories are present in very few instances.

Considering this evidence, the choice becomes evident. Opting for the broader categorization provided by the event type helps mitigate the risk of overfitting the model. The event type, being more balanced and less fragmented than the sub-type, is better suited to capture the overarching trends in the data without being unduly influenced by rare outliers or anomalies.

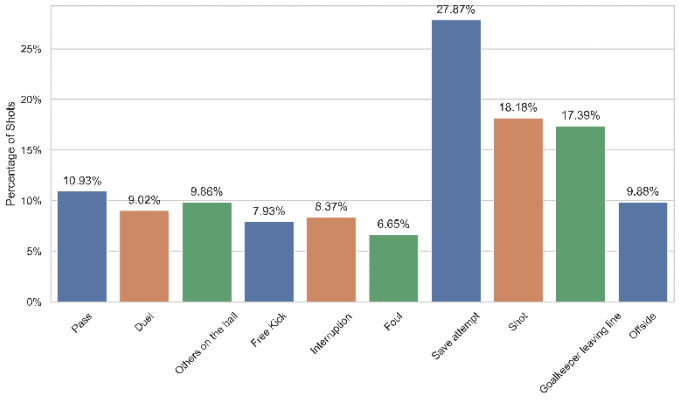

Figure 18: Conversion Rate by Previous Event Type

Fig. 18 illustrates the conversion rates based on the previous event type. Opting for this broader categorization over the more granular sub-types achieves the ideal balance: it provides valuable insights while minimizing the risk of overfitting. The varied conversion rates across different event types further attest to the significance of this choice in our analysis.

Impact of Key Passes:

An intriguing aspect discovered during the analysis pertains to the influence of a 'Key Pass' on shots. In football lexicon, a 'Key Pass' typically refers to a pass that has the potential to directly lead to a goal-scoring opportunity for the receiving player. It is a decisive and impactful pass that breaks through the opponent's defense, creating a clear chance for the attacking team. Its presence or absence can be a potential indicator of the quality or context of the buildup play leading to the shot.

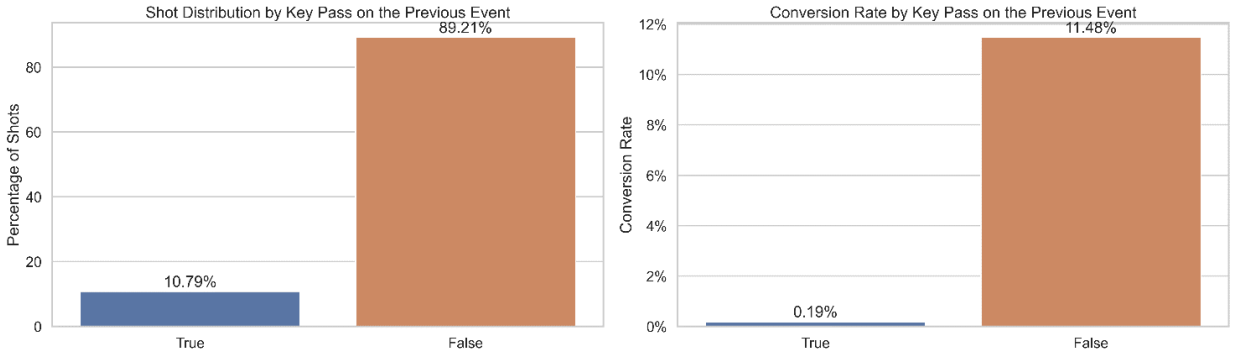

Figure 19: Shot Distribution and Conversion Rate by Key Pass

Fig. 19 provides insights into the frequency distribution and conversion rates associated with the presence of a Key Pass. At a glance, the statistics seem contrary to conventional football understanding: about 89% of shots weren't preceded by a Key Pass, while only around 11% had a Key Pass leading up to them. More confounding is the conversion rate: a mere 0.19% when a Key Pass is present versus a substantial 11% in its absence.

These statistics lead to a fundamental observation about the dataset's nature. In football, it's common for significant passes leading to goals to be labeled as 'assists'. It's plausible that in the dataset, passes labeled as assists aren't concurrently tagged as key passes. This separation might be the reason for the notably low conversion rate following key passes: the truly impactful passes (those that lead to goals) might predominantly be categorized as assists, leaving the 'Key Pass' tag for passes that lead to shots but not necessarily goals.

This observation suggests a possible inconsistency in the dataset. If passes labeled as assists are not also tagged as key passes, it might unintentionally give away the result of the shot. Essentially, using this feature could unintentionally "spoil" the outcome, leading to a model that is misleadingly accurate.

Given this revelation and to ensure the integrity of subsequent analyses, the decision is clear: the 'Key Pass' variable will not be included in the modeling process.

3.5 Associative Analysis of Predictors

To fully understand the relationships within our dataset, it's essential to investigate how the different predictors interact. We'll begin by examining the correlation matrix for continuous variables, providing insight into their linear relationships.

Figure 20: Correlation Matrix

The correlation matrix depicted in Fig. 20 reveals no immediate causes for concern.

Most of the correlations are minor, suggesting that there's no significant linear relationship between most pairs of continuous variables.

The most pronounced correlation appears between the 'previous event angle' and 'previous event distance', registering at 0.45. While this is the highest value in the matrix, it's still not at a level that would typically raise alarm regarding multicollinearity or redundancy.

To understand the relationships between categorical variables instead, two primary metrics are considered:

- Cramér's V [10]: This statistic measures the strength of association between two categorical variables. Its values range from 0 to 1. A value closer to 0 suggests a weak association, while a value closer to 1 indicates a strong association.
- P-values: This is a statistical metric that helps determine the significance of the observed association. Generally, a P-value below 0.05 is considered statistically significant, implying that the observed association is unlikely due to random chance.

Variable 1	Variable 2	P-value	Cramér's V
lastevent_name	prev_Accurate_tag	0.000	0.395
lastevent_name	prev_Won_tag	0.000	0.345
body part	foot	0.000	0.122
situation	lastevent_name	0.000	0.122
prev_Accurate_tag	prev_Won_tag	0.000	0.09
body part	Opportunity	0.000	0.034
body part	lastevent_name	0.000	0.026
situation	body part	0.000	0.023
body part	prev_Won_tag	0.000	0.015

Figure 21: Associations Between Event Categorical Features

In Fig 21, the associations between various categorical variables are laid out, as quantified by Cramér's V and P-values. The table captures a subset of the data — specifically, those associations that have a Cramér's V exceeding 0.01, representing the stronger relationships in our dataset.

Across the board, the p-values are exceedingly low, which suggests that the associations shown in the table are statistically significant and not mere coincidences.

The previous event type demonstrates relatively strong associations with two specific tags: 'Won' and 'Accurate'. This is because they are the predominant tags associated with the most common categories of the previous event type — 'Pass' and 'Duel', respectively.

Beyond the aforementioned strong associations, the table reveals a couple of moderate associations and several weaker ones.

To conclude, the observed strong associations deserve careful consideration during model development. While fitting our predictive model, it will be crucial to evaluate its performance with and without these associated features to ascertain their individual and combined predictive powers. Nevertheless, based on the current analysis, there aren't any associations that raise immediate concerns about multicollinearity or the potential to inappropriately influence the model.

4. Predictive Models

After a thorough exploration of our data, we've gained a clear picture of its many aspects. Now, we're set to tackle the core of this study: building predictive models for Expected Goals.

To ensure a robust evaluation of our models, the dataset was divided into a training set and a test set, using an 80-20 split. Such a division balances the need for ample training data while reserving a significant portion for unbiased model evaluation.

The decision to abstain from resampling techniques was made consciously. The primary objective of this research is not merely to identify as many goals as possible, but rather to assign a meaningful metric that genuinely reflects the scoring potential of a given shot. In this context, resampling might distort the natural balance and frequencies of the data. By preserving the original distribution, the model is better aligned with real-world scenarios, ensuring that the xG metric offers a fair evaluation of a player's chance.

4.1 Metrics

To evaluate the performance of our predictive models, it's essential to choose the right metrics. For this study, we've selected three primary evaluation metrics:

- **ROC AUC Score [17]:** The Receiver Operating Characteristic Area Under the Curve (ROC AUC) measures the model's ability to differentiate between the two classes. An AUC score of 1 indicates perfect predictive power, while a score of 0.5 suggests that the model's predictions are no better than random guessing. It's robust against imbalanced datasets because it evaluates the model's performance across all possible thresholds. However, it might sometimes give an overly optimistic view of the model's performance, especially if one class (Goal in this case) is very rare.
- **Precision-Recall AUC [11]:** Precision and Recall are two critical metrics for evaluating the performance of models, especially when dealing with imbalanced datasets. Precision measures the accuracy of the positive predictions, while Recall (or Sensitivity) measures the percentage of actual positives that were identified correctly. The Precision-Recall curve plots Recall on the x-axis and Precision on the y-axis for different thresholds. The area under this curve (AUC) gives us a single value that captures the overall performance of the model. In scenarios having a significant class imbalance like this, the Precision-Recall AUC often provides a more realistic evaluation of the model's performance compared to the ROC AUC.
- **Log Loss [18]:** Quantifies the accuracy of a classifier by penalizing false classifications. Lower log loss indicates better predictive accuracy.
  Given that our predictive models aim to estimate a probability of the binary outcome (Goal vs. Not Goal), log loss is a proper metric as it directly evaluates the quality of these probability predictions.

By employing these metrics, we aim to comprehensively evaluate the model's performance, ensuring that it's both robust and accurate in predicting Expected Goals.

4.2 Models

To predict Expected Goals, a selection of machine learning models was employed:

1. Logistic Regression [12]
2. Random Forest [13]
3. Gradient Boosting Machine (GBM) [14]
4. Support Vector Machines (SVM) [15]
5. Neural Networks [16]

Each model underwent meticulous fine-tuning for optimal performance. Hyperparameters were adjusted, and feature selection was applied when beneficial. The table in Fig. 22 provide a comparative analysis of their performance metrics, offering insights into their effectiveness for this dataset.

	Logistic Regression	Random Forest	GBM	SVM	Neural Networks
ROC AUC Score	0.825	0.820	0.827	0.760	0.810
Precision-Recall AUC	0.364	0.368	0.377	0.250	0.332
Log Loss	0.261	0.262	0.259	0.290	0.273

Figure 22: Models Performances

The **Gradient Boosting Machine (GBM)** emerges as the standout model, excelling in all three metrics: with a ROC AUC Score of 0.827, a Precision-Recall AUC of 0.377, and a Log Loss of 0.259.

Other models offer also competitive performances, closely trailing GBM. Support Vector Machines (SVM) instead have less impressive metrics, particularly its Precision-Recall AUC of 0.250, suggesting it struggles with the imbalanced nature of the dataset.

Evaluating our best model, the GBM, the results are promising. The ROC AUC Score suggests a relatively strong discriminative power, The Precision-Recall AUC approaching 0.38 in an imbalanced dataset highlights its capability in capturing the Goal outcomes. The Log Loss of 0.259 further attests to its accuracy in delivering probability estimations.

Following the performance evaluation, we turn our attention to the calibration of our best-performing model.

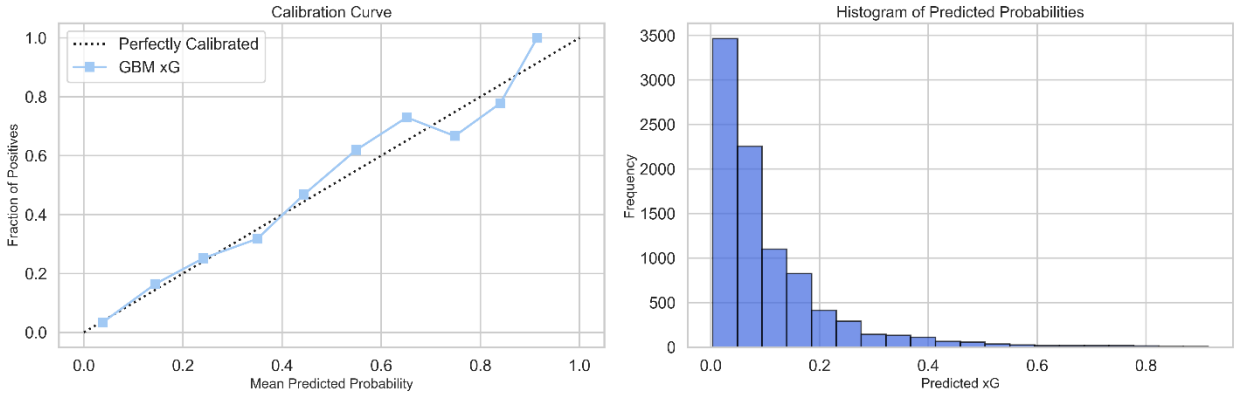

Figure 23: Calibration and Histogram of Predicted xG

The calibration plot in Fig. 23 offers a visual assessment of our model's predictions against actual outcomes. Impressively, for predictions of xG below 0.5 — which constitute the vast majority of our predictions — the GBM's calibration is nearly impeccable, closely aligning with the ideal calibration. As we shift to higher xG predictions, there's a slight deviation from the optimal calibration curve. Yet, given the limited instances of such high xG predictions shown in Fig. 23, this divergence is relatively minor, suggesting that the model's calibration remains robust across the spectrum of predictions.

In our pursuit of understanding the underlying mechanisms of our best model feature importances are shown.

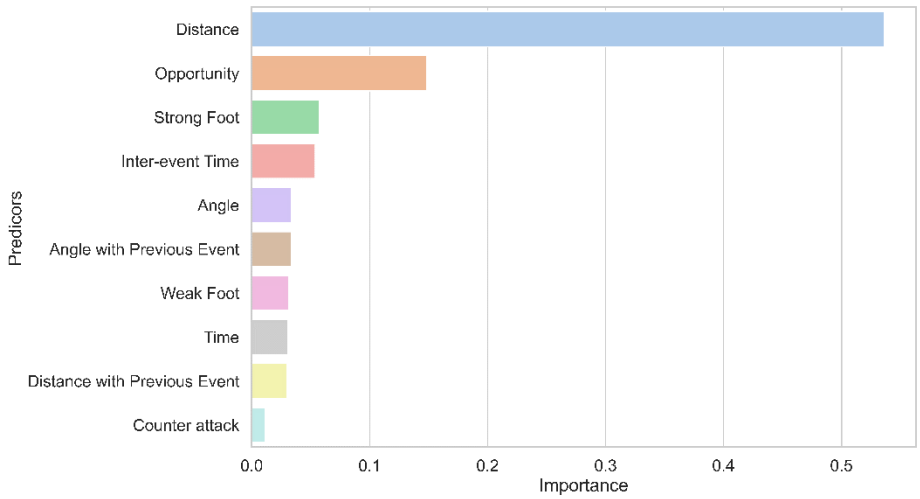

Figure 24: Feature Importance Barplot

The visual representation in Fig. 24 shows coefficents of the importance given by GBM to its top 10 most important features. Distance shows a clear dominance, its importance value exceeding 0.5 define its pivotal role in determining Expected Goals. Opportunity also stands out with a substantial importance score of over 0.1. These two features collectively play a considerable role in the model's decision-making process.

All the remaining features have importance values below 0.1. This highlights the comparatively lesser impact they have in shaping the model's predictions.

4.3 Benchmarking Against Literature

In the subsequent table, we compare the performance of our models against those documented in academic papers [4], [5], [6], [7], [8], [9]. The metrics of comparison are the ROC AUC Score and the Log Loss. Regrettably, the models from these studies did not employ the Precision-Recall AUC, limiting our comparative analysis to the two metrics.

Paper	Model	AUC	Log Loss
Eggels et al. (2016) [4]	Logistic Regression	0.697	-
	Random Forest	0.814	-
	Gradient Boosting (adaboost)	0.670	-
	Neural Network	-	-
Pardo (2020) [5]	Logistic Regression	-	**0.256**
	Random Forest	-	-
	Gradient Boosting (xgboost)	-	0.254
	Neural Network	-	**0.255**
Anzer and Bauer (2021) [6]	Logistic Regression	0.807	-
	Random Forest	0.794	-
	Gradient Boosting	0.822	-
	Neural Network	-	-
Haaren (2021) [7]	Logistic Regression	-	-
	Random Forest	-	-
	Gradient Boosting	0.793	0.288
	Neural Network	-	-
Cavus et al. (2021) [8]	Logistic Regression	-	-
	Random Forest	**0.975**	**0.173**
	Gradient Boosting (catboost)	0.823	0.261
	Neural Network	-	-
Mead J et al (2023) [9]	Logistic Regression	-	0.286
	Random Forest	-	0.290
	Gradient Boosting (xgboost)	0.800	0.282
	Neural Network	-	0.283
Our models	Logistic Regression	**0.825**	0.261
	Random Forest	0.820	0.262
	Gradient Boosting	**0.827**	**0.259**
	Neural Network	**0.810**	0.273

*In bold are highlighted the best metrics for each model

Figure 25: Models Comparison with Literature

The table in Fig. 25 provides a comparative snapshot of our models against those documented in a selection of prominent academic papers. Some initial observations from the table are:

Among the six papers referenced, [4], [5], and [9] incorporate measures of player or team ability in their models. This introduces a potential bias, as we discussed at the outset of our thesis, especially when xG is used to evaluate player or team performances.

On the contrary, papers [6], [7], and [8] avoid such potential pitfalls by not relying on measures of player and team ability.

Best performing models:

- Our GBM model stands out with top-tier performance, showing the highest scores in both AUC and Log Loss.
- Our logistic regression model achieves the highest AUC, though it's slightly edged out in Log Loss by [5], which, as noted, incorporates player ability metrics. The unreported AUC for [5] is likely to be higher, given its superior Log Loss.
- The best Random Forest model is presented by [8], excelling in both metrics.
- For Neural Networks, our model again leads in AUC but is marginally outperformed in Log Loss by [5]. Given the Log Loss results, [5]'s AUC is presumably higher.

A striking revelation is the performance of the Random Forest model by Cavus et al. [8]. It not only surpasses our models but also all the others, without the inclusion of player/team predictors.

Our models have demonstrated strong performance, even surpassing some models that leverage player and team predictors. This underscores the efficacy of our approach and the quality of our dataset.

The standout performance of the Random Forest model by Cavus et al. [8] is both impressive and puzzling. Their model relies on basic features, many of which overlap with ours, yet it achieved superior results. Despite our extensive feature engineering, their simpler approach delivered better benchmarks. The specifics of their model fitting process aren't detailed, making it challenging to pinpoint the exact reasons for its exceptional performance.

5. Conclusion

The realm of football analytics has witnessed the meteoric rise of the Expected Goals (xG) metric, offering a lens to evaluate and understand the game's dynamics beyond mere goal counts. Yet, a critical bottleneck has persisted. As outlined in the initial chapters, many extant xG models, in their quest for predictive accuracy, inadvertently introduce bias by incorporating player and team attributes. Such an approach, while possibly elevating

prediction rates, undermines the xG's core value: offering an objective, unbiased measure of shot quality. When the xG metric is muddled with influences of who took the shot or which team they played for, its efficacy as a neutral evaluator wanes.

Our research was galvanized by this very challenge. The aim was clear: to create an xG model rooted solely in situational features, excluding potentially biasing attributes. The results, as revealed in the subsequent chapters, have been both heartening and revelatory. We've not only demonstrated the viability of such an unbiased approach but, in many instances, our models have matched or even eclipsed the performance of their conventional, more biased counterparts. This accomplishment underscores the robustness of situational features in capturing the essence of shot quality, without the need for potentially confounding player or team metrics.

What does this mean for football analytics? First and foremost, it reinforces the potential of xG as a transparent, impartial tool for player, team, and match assessments. Such an unbiased metric opens avenues for deeper, more nuanced analyses, allowing stakeholders to dissect performances without the shadow of inherent biases. Furthermore, by emphasizing the shot's situational context over individual attributes, the findings set a precedent for future research, nudging the football analytics community towards more bias-free model constructions.

Imagine the profound implications: evaluations of players based not on their reputation, but on their ability to carve out genuine goal-scoring opportunities; assessments of teams that focus on the quality of chances they create or concede, rather than just the star power in their ranks. In essence, we're championing a paradigm shift, nudging football analytics towards a more objective, situational lens.

As we march forward in the analytical age of sports, let this be our guiding light: to understand the game in its purest form, celebrating every nuance, every strategy, and every unrewarded hero. The Beautiful Game deserves nothing less.

References

1. Soccerment advanced metrics. Available at: https://soccerment.com/soccerments-advanced-metrics/.
2. Pappalardo, Luca; Massucco, Emanuele (2019). Soccer match event dataset. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.4415000.
3. Wyscout API Documentation. Available at: https://apidocs.wyscout.com/#section/Data-glossary-and-definitions.
4. Pardo, M. (2020). Creating a model for expected goals in football using qualitative player information. Master's thesis, Universitat Politecnica de Catalunya.
5. Eggels, H., van Elk, R., & Pechenizkiy, M. (2016). Explaining soccer match outcomes with goal scoring opportunities predictive analytics. In Proceedings of the Workshop on Machine Learning and Data Mining for Sports Analytics 2016 co-located with the 2016 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2016). CEUR-WS.org. (CEUR Workshop Proceedings).
6. Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). Frontiers in Sports and Active Living, 3, 624475. https://doi.org/10.3389/fspor.2021.624475. PMID: 33889843; PMCID: PMC8056301.
7. Haaren, J. V. (2021). Why would I trust your numbers? On the explainability of expected values in soccer. arXiv preprint, arXiv:2105.13778.
8. Cavus, M., & Biecek, P. (2022). Explainable expected goal models for performance analysis in football analytics. arXiv:2206.07212.
9. Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. PLOS ONE, 18(4), e0282295. https://doi.org/10.1371/journal.pone.0282295.
10. Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.
11. Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 233–240.
12. Cox, D. R. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological), 20(2), 215–242.
13. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
14. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189-1232.

15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
16. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.
17. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.
18. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press