Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier

Milestone Report

Problem Statement

European soccer is big business. The 'big five' leagues, made up by England, Germany, Italy, Spain and France posted $17.4 billion USD revenue for the 2017/18 season, an increase of 6% from the previous year. Increasing team performance can have a direct impact on revenue by increasing viewership and increasing prize money awarded at the end of the season. Currently analysis is performed manually and teams spend a lot of effort reviewing tapes to identify winning tactics and strategy. Automating this process would provide value to teams by increasing the quality of the analysis and even being able to predict the impact changing tactics or scouting a new player will have on their performance.

The data for this project was generated by Wyscout and was generated by reviewing tape and cataloging all of the ball touches throughout the game. To start, the players starting positions on the field are documented and then every interaction with the ball is tagged as an event including the position on the field and the type of event pass, duel, shot, etc. Events leading to a shot on the goal are also classified to indicate whether or not a goal was scored and the position of the shot on the goal. The data is organized into JSON files which are easily parsed using the Pandas read_json function. Preliminary analysis shows the data is free of null values and the datatypes are sensible for each factor so it is anticipated that the data will be relatively easy to use for the scope of this project

The goal of this project is to be able to use machine learning to identify and classify patterns and formations and then assess how these patterns, formations, the qualities of players in these formations and their interactions with each other predict success in terms of how likely it is for a given team to win the match. Ultimately the aim is that this can be used by team managers/owners to determine how changing formation, substituting a player, or perhaps recruiting new talent will impact their odds of having a winning season.

Data Wrangling

The data set I selected for this project was relatively clean. The main data wrangling that was needed was in restructuring my data. Some features, for example, contained lists of dictionaries which needed to be unpacked into their own individual columns to allow for better data analysis. In addition, there were some missing values which needed to be handled and

some filtering was applied to reduce the scale of the dataset. All data was imported from JSON files using the Pandas.read_json().

The data set contains various data from the 2017 season in the English, German, Italian, French soccer leagues as well as events from the World Cup and the European Championship. For all of these competitions, there is match data, player data, team data, and event data. The analysis will likely focus on the event data where each observation is an interaction with the soccer ball. The events dataset for the English Premier League contains 643,032 observations which seems sufficiently large for the scope of this project and so as part of the data wrangling effort the data was filtered down to the Premier League only. The Premier League was selected because it generates the highest revenue of all the European leagues.[1]

In reviewing the teams dataset, the country name and ID are nested in a dictionary in the 'area' column. In order to make this more accessible for analysis these values were extracted and stored in their own columns by utilizing the map function in conjunction with a lambda function:

teams['countryId'] = teams.area.map(lambda v:v['id'])

The 'countryId' was then used to filter out only the Premier League teams, followed by the 'type' which removed the national teams leaving only the club teams. The data was then sorted by teamId and then re-indexed. Similar challenges were encountered in the players data set and the same map/lambda function combo was used to extract player positions.

While working through the player and events datasets I also encountered some missing values. In the players dataset, this included values for players weights, heights, foot (left/right/both), and currentTeamId. The weights/heights had some values listed as 0 and the foot values contained both null and empty strings. All of these were replaced with numpy.nan using the Pandas replace() function. This was done prior to filtering for the Premier League to allow for flexibility in expanding the analysis. After filtering for the Premier League there was only one player with NaN values listed for all three of these attributes and so to correct this, I manually looked up these values with a quick Google search and updated them in the dataframe manually. Lastly, when reading in the json file I had to set the encoding parameter to

[1] Kidd, Robert. "Powerhouse Premier League Helps European Soccer Reach Record Revenues." Forbes.com. Forbes, 30 May, 2019. Web. <https://www.forbes.com/sites/robertkidd/2019/05/30/powerhouse-premier-league-helps-european-soccer-reach-record-revenues/#4b836d2923ef>

'unicode-escape' to account for special characters in the players' names. When saving the wrangled data in csv format, I opted to use the utf-8 encoding instead as this is the pandas default and will simplify importing in the future.

In the matches and events datasets again there some of the features included dictionary objects which needed to be unpacked. In these cases, however, I utilized a for loop with a list comprehension to build a list of the tag from the dictionaries. The list comprehension enabled me to easily handle the varying lengths of the tag dictionaries. I would like to revisit this if time allows, to come up with a more elegant and efficient solution. The for loop does not take advantage of the vectorized operations but after spending a while trying to figure out other solutions I decided to move forward with the most basic method.

```
# create new tags column with list instead of dictionary
tag_list = []
for index, row in events.iterrows()
    tag_list.append([d['id'] for d in row['tags']])
```

In addition, I chained the .values.tolist() methods to the events.positions series to unpack the position of each event that was recorded as a list of two dictionaries, such as [{'y': 49, 'x': 49}, {'y': 78, 'x': 31}]. The first representing the starting position of the event and the second represents the end position of the event.

```
events[['start_pos', 'end_pos']] = pd.DataFrame(events.positions.values.tolist(),
index=events.index)
```

By repeating this two more times I was able to create 4 columns, one for each x/y position at either the start or end of each event which will allow me to easily plot and analyze how the ball moves around the field and the effect this has on a team's performance. Some of the events did not have an end position because they were interrupted by a foul or protest. These were removed for simplicity with the assumption that they are not performance related and are not relevant to the analysis.

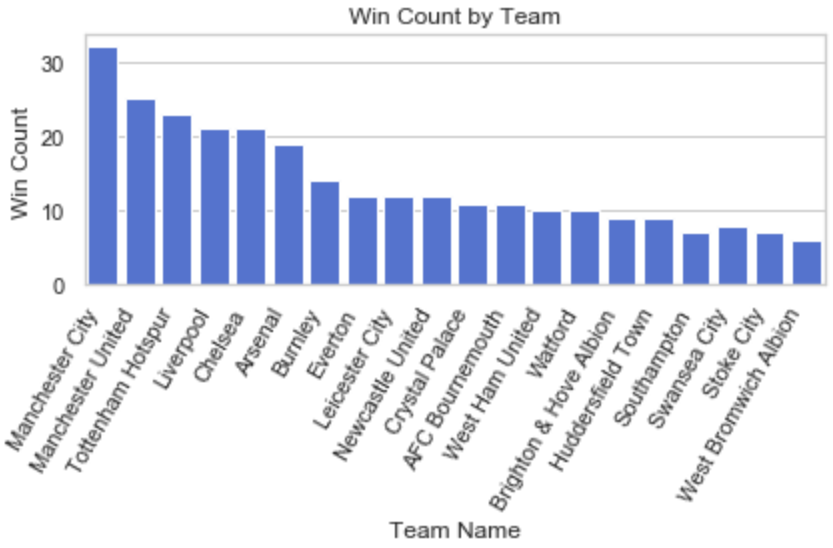
I did not encounter any outliers while working through the dataset.

Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier

Exploratory Data Analysis

Let's first explore the team standings at the close of the season. The figure below shows the total number of wins and is sorted based on team standing. We can see clearly that the number of wins has a strong correlation with where the team finished in the standings, though there are some instances like Southampton where they placed ahead of teams who actually won more matches.

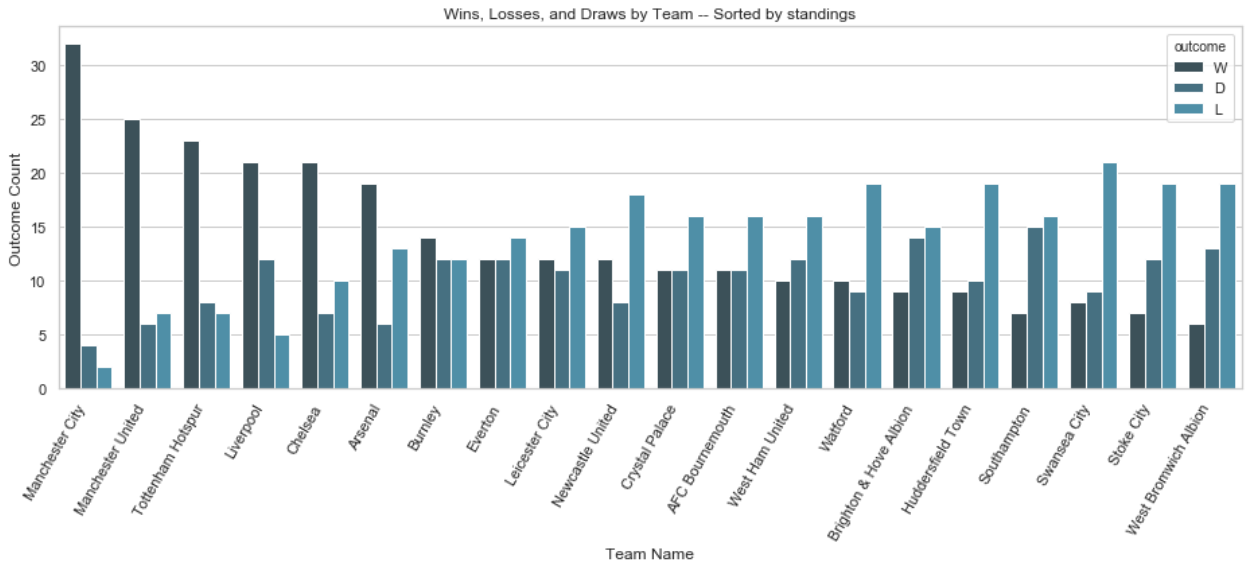


The plot below shows draws and losses in addition to the wins shown in the first plot. We can see that to the left of the plot the wins greatly outweigh draws and losses. Toward the middle of the plot, the three outcomes are all balanced, and toward the right the number of losses starts to outweigh wins and draws.

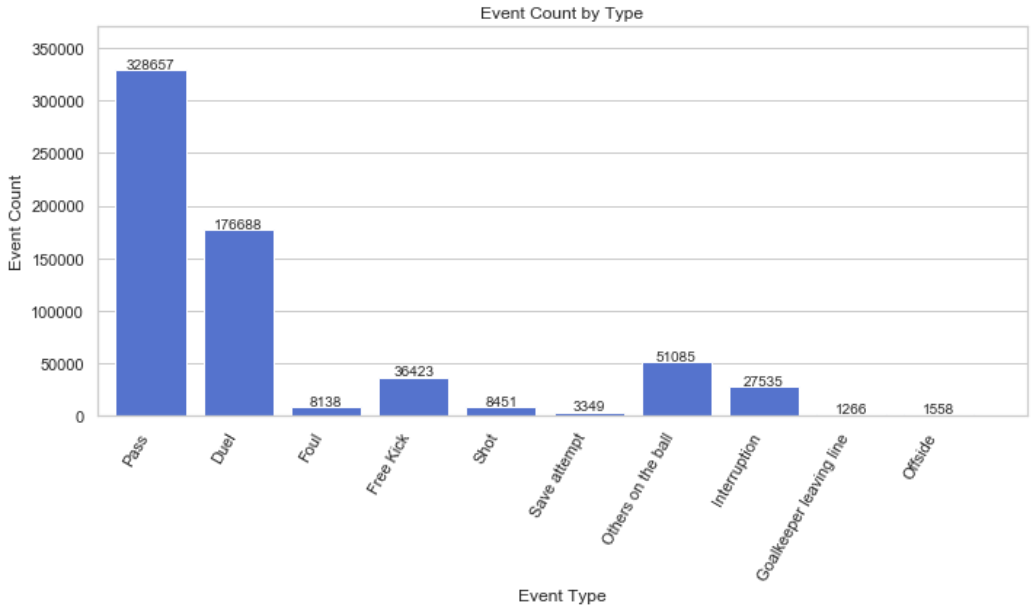
This is sensible considering the points system used in the premier league which is used to determine team standings. A win is awarded 3 points, a draw 1, and a loss 0 points. Ties in points are decided by goal difference and then by number of goals scored. In this case goal difference is the number of goals a team scored minus the number of goals other teams scored against them.

Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier

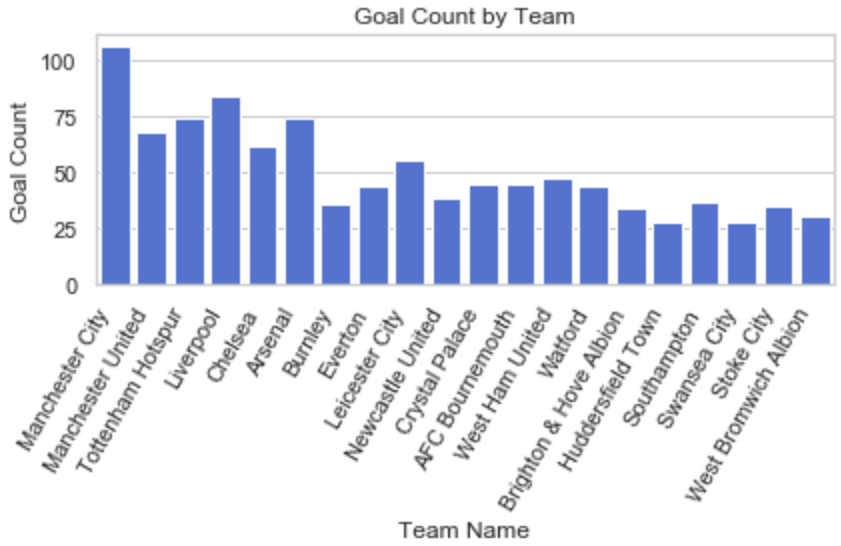


Now that we've looked at the top level performance of each team, let's explore the events themselves. To start, let's see which types of events are most common. The plot below shows that passes are the most common type of event and that shots themselves are only the sixth most common type of event.

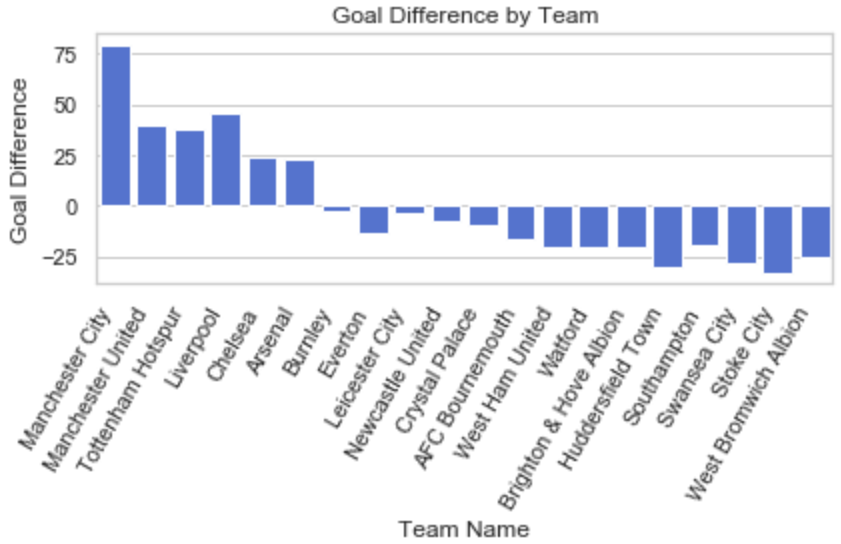


Next let's see how the total number of goals scored in a season relates to team standing. The below plot shows that there is more to winning than just the number of goals scored. Liverpool, for example, scored the second most number of goals but still only finished fourth in

the season. This could be because they scored many goals in a handful of wins and/or because they finished more games in a tie. Let's take a closer look.



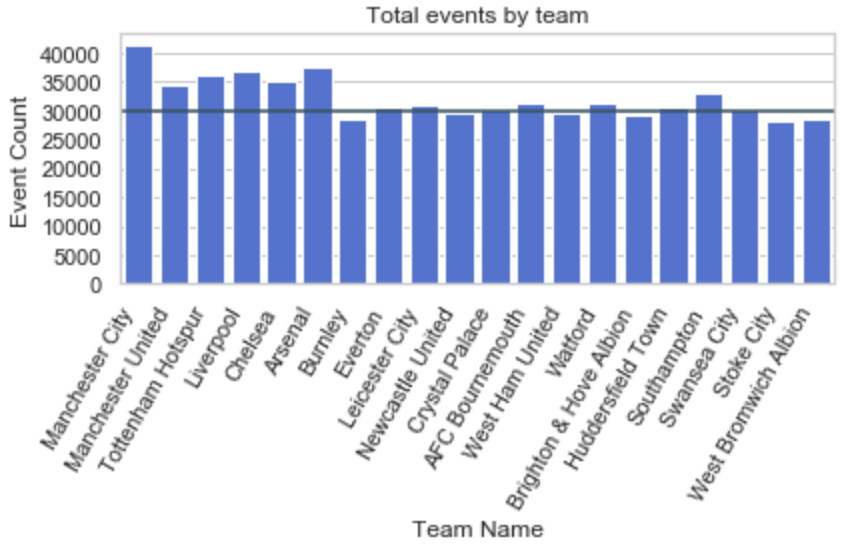
The goal difference is calculated by subtracting the number of goals scored against a team by the number of goals that team scored themselves. When we look at a plot of these values sorted by standings we see that the top 6 teams stand out from the other teams in that they have a positive goal difference. This means they've scored more goals than have been scored against them.



Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier

Now let's consider the total number of events throughout the season by team. Below we see that the top six teams were involved in more events overall than the rest of the field, with the exception of Southampton. This seems to suggest that these teams are dominant and maintain possession of the ball more than the other teams.

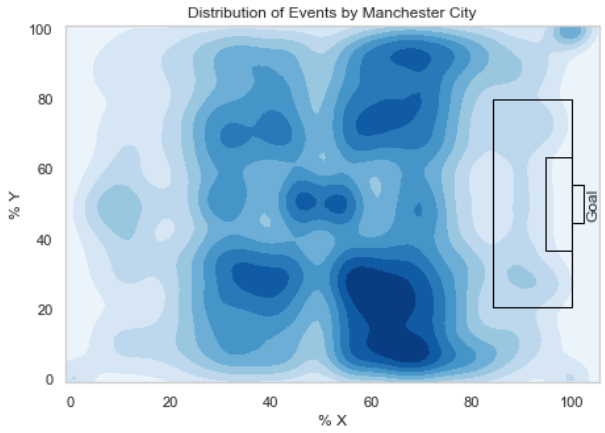


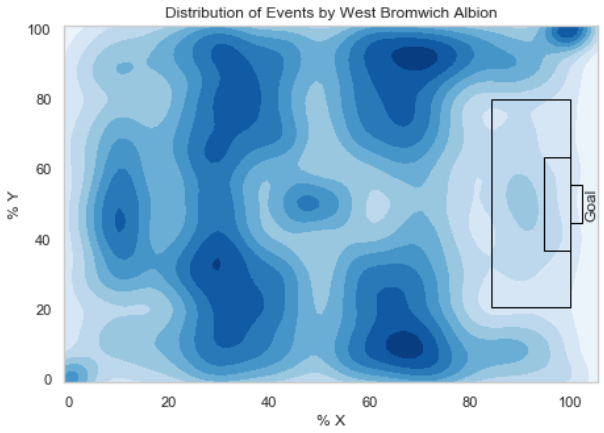
Rather than just considering the total count, let's explore the distribution of events across the field. We'll start by comparing the distribution of events across the field for Manchester City, who finished first, and West Bromwich Albion, who finished last.

There is some variation in the exact size of the playing field across the premier league. As such, all x and y positions are reported as percentages of the field. The x-y origin is set as the lower left hand corner and all play occurs with the attack heading toward the right. This is a necessary form of normalization because teams switch sides half way through the game and change directions. By making the attack always play to the right it assures that all plays to the right of the midway point are offensive and all plays to the left of the midway point are defensive.

Predicting success in european soccer based on tactical spacio-temporal data

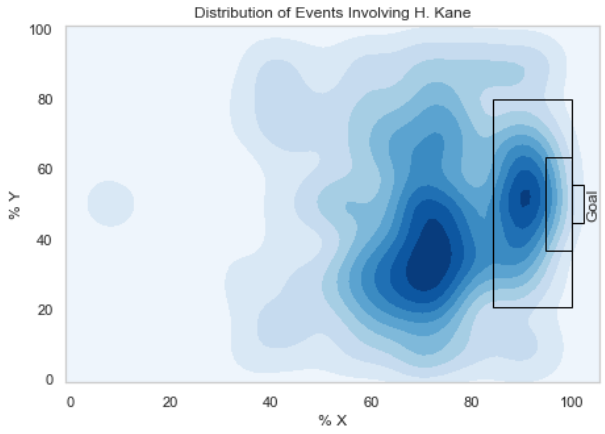
by Ryan Lussier

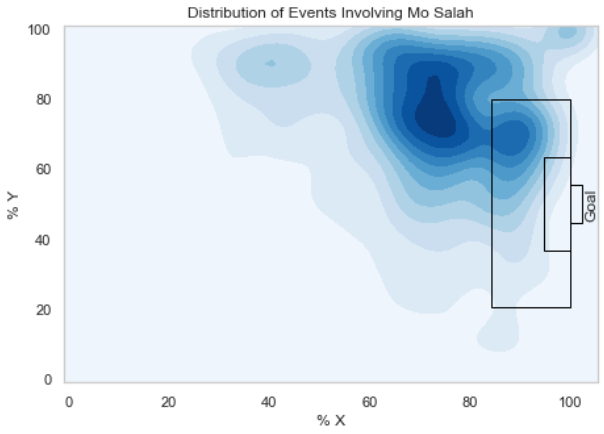




Here we see that it appears that West Bromwich Albion (WBA) has a higher density of events in the backfield where they are on the defence whereas Manchester city has a bias toward the attack. We especially see this is true in the defending goal box where WBA has more occurences of their total events whereas Manchester city has relatively few. We can also see that WBA has a high density in the top right corner which indicates a large number of corner kicks. These are awarded to a team when the ball goes out of bounds last touched by their opponent at the opponent's own end of the field. In this case it may be an indication of WBA not being able to convert potential offensive opportunities to shots but rather having the ball stolen and/or deflected out of bounds.

The previous plots show trends as it relates to entire teams. Let's take a closer look and compare two individual players. In this case let's compare Harry Kane and Mohamed Salah -- two leading forwards in the league.



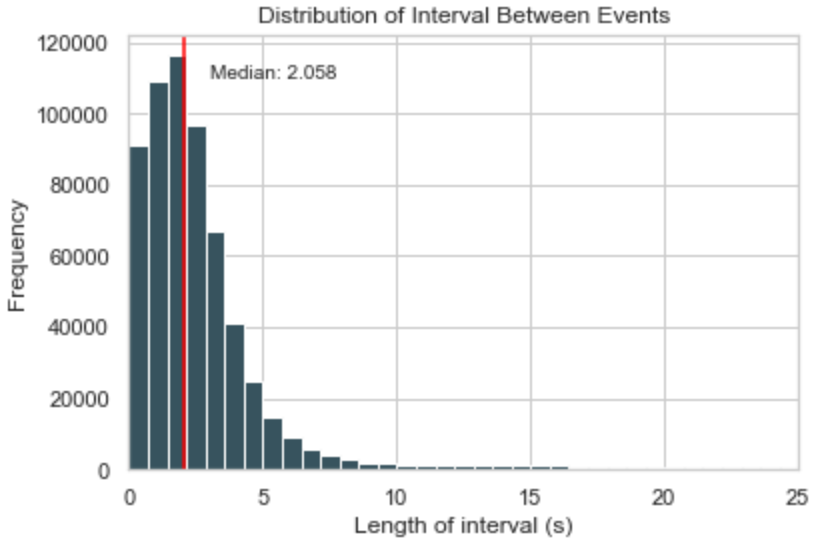


Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier

It is clear that Mo Salah prefers to play toward the top end of the pitch and the top left corner of the goal box. This is very different from Harry Kane who appears to play a more centered game and has a larger presence in the center of the goal box. For reference, Mo Salah led the league with 32 goals this season while Harry Kane scored 30 goals.

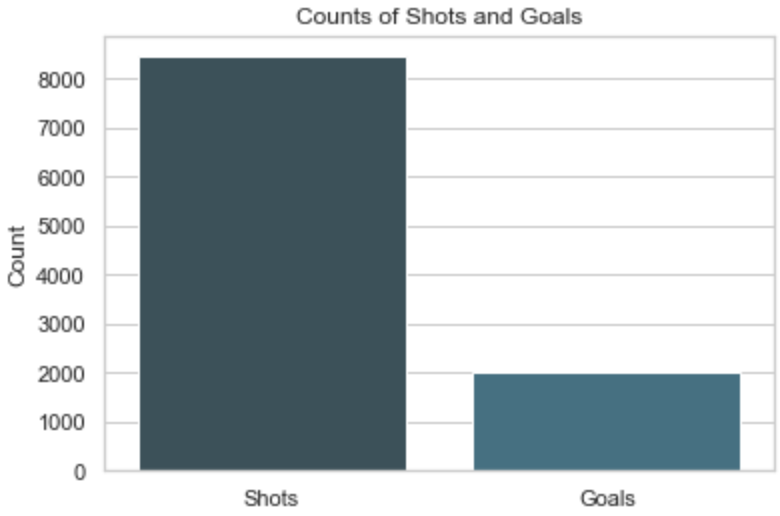
One hypothesis I have is that goals are more likely when the passes leading to them occur rapidly. To start exploring this, let's look at the distribution of intervals between events. This is simply the difference in the time each event was logged. The histogram below shows a gamma distribution with a median value of 2.058 seconds. This shows that most events are relatively short to begin with gives context for what length of interval should be considered rapid.



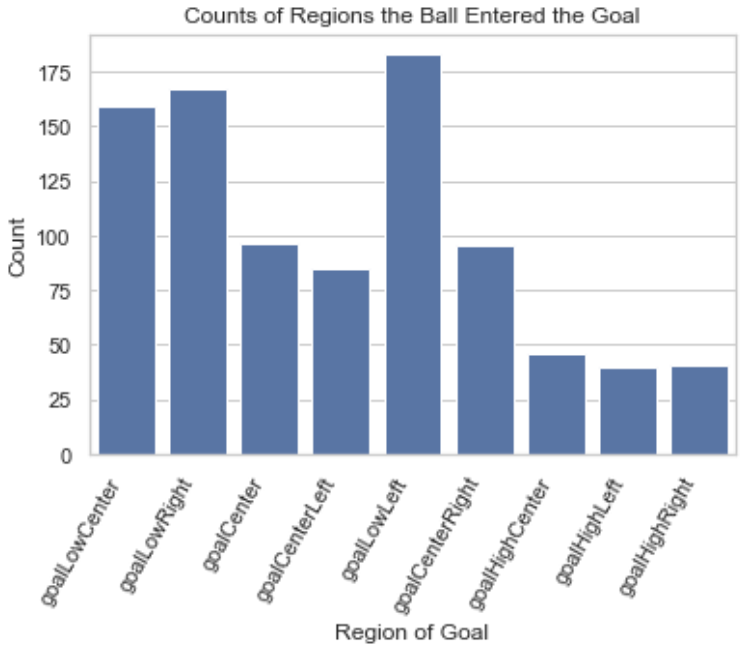
Previously we explored how goal count and goal difference correlated with the final standing of each team. Let's take a closer look at the details of the goal events themselves. Below we see that the percentage of goals scored is much smaller than the number of shots attempted. Only about 25% of shots result in a goal. This unbalance should be taken into consideration later when building predictive models.

Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier



One potentially interesting observation is the region of the goal the ball enters as it breaks the plane of the goal. Picture dividing the goal into bins arranged in a 3x3 grid and labeling it from the attacking player's position. Vertically we label the bins low, center and high. Horizontally we label the bins left, center, and right. We then count the number of goals that enter each bin.



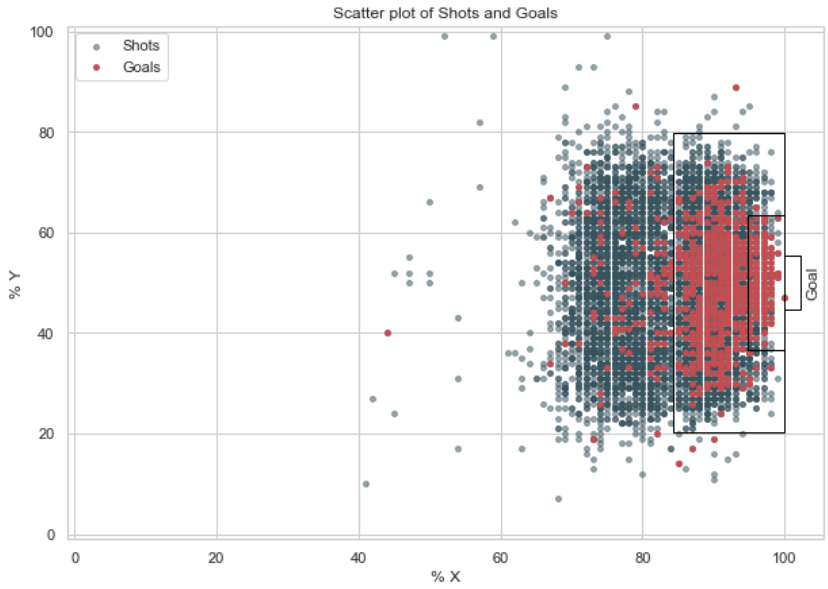
Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier

Here we see that most goals are scored in the lower third of the goal and the least goals are scored in the upper third of the goal. The left-to-right distribution is fairly balanced. The lower left bin has slightly more goals than the other bins in the lower portion of the goal.

It is unclear if these trends are simply a result of the number of shots taken at each region. Further investigation comparing the percentage of misses in the various regions around the goal will provide more insight.

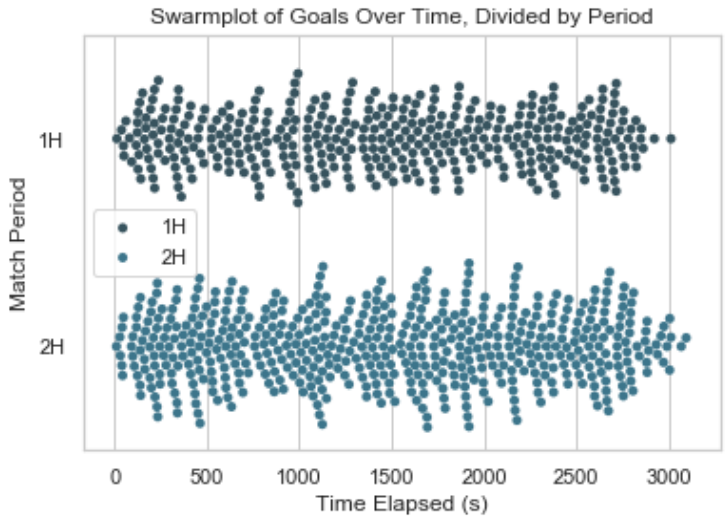
The plot below shows the location of every shot and every goal. It is clear that a higher percentage of goals are made when the shots are taken closer to the goal. This is sensible both because the player shooting the ball should have an easier time aiming and hitting their target from a shorter distance. In addition, the goal keeper and defensive players will have less time to react to the ball before it reaches the goal line. Another observation is that the shots and goals are centered vertically relative to the goal. In fact most shots are taken within the y bounds of the penalty area.



Lastly, let's explore if the frequency of goals changes over time. Below is a swarm plot showing the timing of each goal scored where the data are grouped by period. It is clear that overall more goals are scored in the second half of the game than in the first. This may be a result of players fatiguing over time and the defence making more mistakes which ultimately result in more goals. Calculating the percentage of goals scored in each period we see that almost 60% of goals are scored in the second period.

Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier



To summarize, it is clear that goal difference is a strong indicator of a team's success overall. A deeper dive showed that x/y position of shot attempts and match period seem to have an impact on scoring a goal. We also saw that more goals are scored in the lower portion of the net which may indicate that the region of the goal the ball is aimed influences the chance of scoring a goal.

Hypothesis Testing

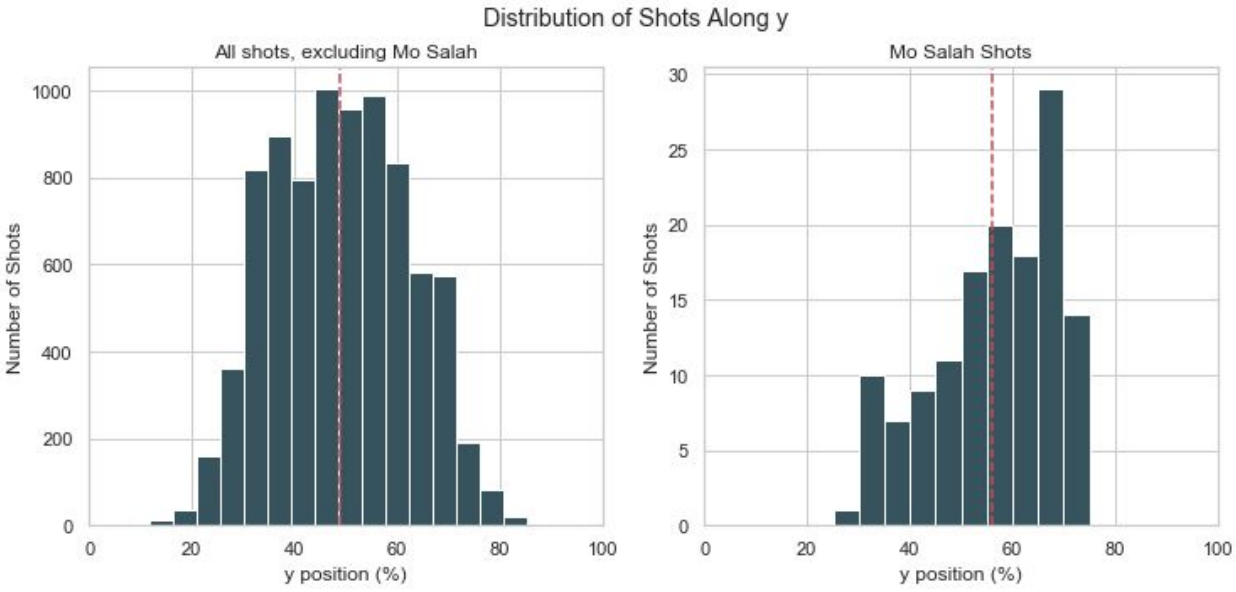
During our exploratory data analysis it appeared that events involving Mo Salah were biased toward the left side of the field. Let's take a closer look and explore the question: is Mo Salah's mean y position when he shoots the ball different from everyone else? To evaluate this hypothesis we will assume Mo Salah is independent of the rest of the field and perform an independent t-test. To start, let's establish our null and alternative hypotheses.

H0: Mo Salah shoots the ball from the same y position as everyone else.
H1: Mo Salah shoots the ball from a different y position than everyone else

Before we get to the t-test, however, let's take a look at the distributions.

Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier



In the histograms above, 0 on the x-axis represents the right sideline if you are facing the goal and 100 represents the left sideline. The positions are shown as a percentage of the field width. It appears that the mean of all players excluding Mo Salah is close to 50% while Mo is biased towards the left side of the goal.

Now lets run the t-test to see if this difference is significant. Typically normality is assumed when running a t-test. This assumption seems to hold for all shots, but when looking at Mo exclusively the distribution is not as normally distributed. In this case we will overlook the less than ideal distribution because t-tests are robust to violations of the normality assumption. To read more about this robustness, check out this link.

For the t-test we will utilize the ttest_ind function from scipy.stats which yields the following results.

t statistic: 6.140147247988889

p value: 8.610519188761847e-10

The results of this test show a p value which is essentially 0 which makes a strong case for rejecting the null hypothesis. Let's take this a step further and compare the t-statistic to the critical t score. For this case we'll set our confidence level to 0.95.

critical t score: 1.6450339959007674

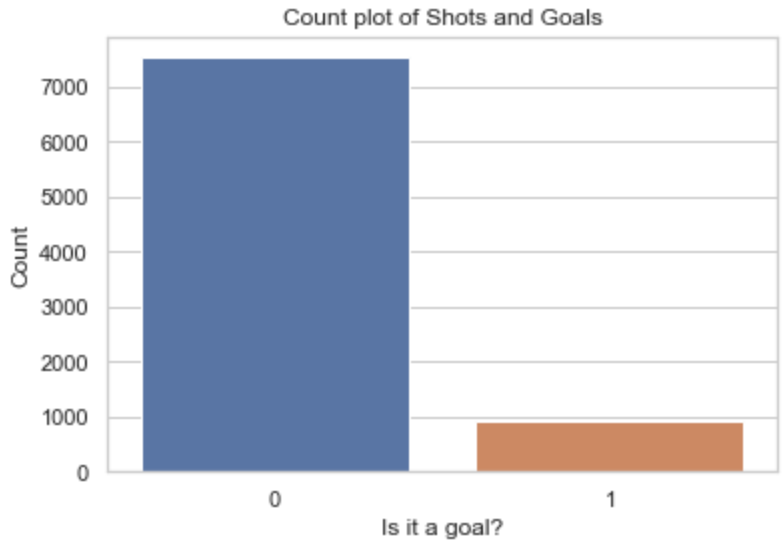
Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier

If we compare the t-statistic with the critical t score we find that 6.14 > 1.65 and we reject the null hypothesis. Thus we can conclude with 95% certainty that on average Mo Salah shoots the soccer ball from a different y position than everyone else. One possible explanation for this may be that Salah prefers his left foot and thus plays to the left side of the goal where it is easier to control the ball and cross back toward the goal or the center of the field.

Basic Logistic Regression

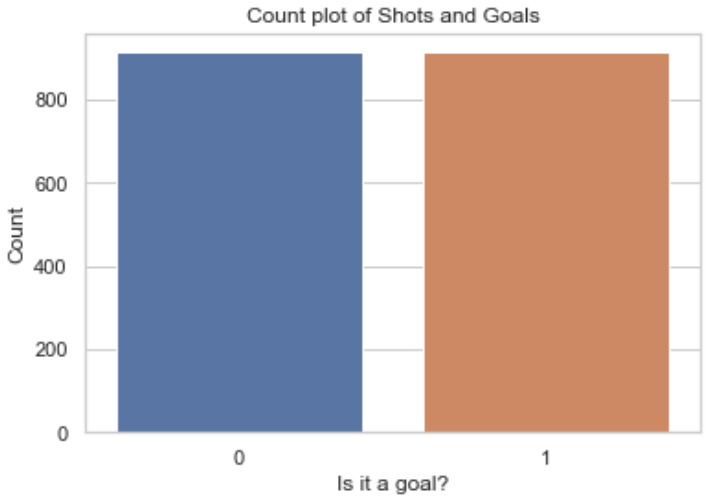
Before investing a large amount of time developing a machine learning model for predicting success, I thought it would be worthwhile to practice implementing a simple model. During EDA we saw that goal count was a decent predictor of team rank so we will take a closer look at predicting goals given a shot event. In this case, I'll use a basic logistic regression using x/y position of a shot as a feature to predict if a goal will be scored.



As you can see the data is very imbalanced, with much fewer goals than misses. To solve this problem for our preliminary model we will simply undersample the missed shots.

Predicting success in european soccer based on tactical spacio-temporal data

by Ryan Lussier



The above count plot shows that the dataset is now balanced. However, the downside of undersampling is that we have just thrown away almost 80% of our data. This will have a negative impact on the model's predictive power and we will explore other options as we make the model more robust.

Moving on, we utilize functions from the sklearn libraries of model_selection, linear_model, and metrics to implement and evaluate this basic model. For the train/test split I opted for an 80/20 split because we discarded so much data during undersampling. The sklearn logreg score for this classifier was 0.68. The confusion matrix below shows the results in more detail. The largest error seems to be coming from misclassification of misses as goals. Overall this is not a bad start for such a small time investment. Going forward I will implement a logistic regression with more features and then explore other models such as random forests and/or sequence classification which may gain insights from the sequence of moves leading up to a shot.

		Predicted Label	
		Miss	Goal
True Label	Miss	116	81
	Goal	35	134