**I. Introduction**

This is a paper that outlines a soccer betting model aimed at beating the generic market-makers for soccer outcomes. The model produces outcomes that allow for bets on moneylines, spreads, totals and player props. I do not have the necessary information to guess at how sportsbooks calculate their soccer odds, but it is well-known that soccer models based solely on ratings systems like Elo and 538’s spi are not likely to [produce profitable outcomes](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5988281/).

However, these simple models miss several factors: Starting lineups, current team manager, substitutions, formations, league factors -- among many others. My goal is to combine as many game-to-game variables to the analysis possible via these ratings to produce the most accurate possible model. In this paper, I will, to the best of my ability, explain the segmentation that I have done, why I think it necessary, and how it works within the context of predicting the outcome of a game.

**II. Methodology**

I collected the last three years of game data via Fotmob (via Opta, I believe) for every possible league where there was data down to the player level. There are several obvious identification variables that are necessary for the model, and some that are specific to the dataset. The variables that are specific are as following:

* Minutes
* Games
* Minutes where XG data is tracked
* Passes
* Shots
* Shots on Goal
* Expected Goals
* Expected Goals on Target
* Expected Assists
* Fotmob Rating

First, I downloaded the data for each day and created a folder for every single player in the database -- putting each player’s game file individually in their corresponding folder. I then segmented each day’s data further by manager, league, formation matchup, and team. I will explain why these are what I considered the most important groups later.

I assigned a rating to each team where a rating was available -- using 538 for club ratings and worldeloratings.net for International Team Rankings. These are the bases for the regression models that create the expected stats that inform the game predictions.

The model is structured by FotMob. I first get the games by league and then take it game-by-game to make specific predictions. All data comes from the leagues in which FotMob has passing-level data. The analysis works in the order as follows:

1. I first get every game that every player on the teams have played in the past three years and combine them into one dataset by the position that player played.
2. From there, I get the lineups from FotMob. There are currently no lineup predictions, though those will come later. FotMob offers probable lineups for popular leagues no more than a day in advance; however, they use the opta lineup data as soon as it is available to display lineups when available. I merge the player stats to the lineups, including raw stats as well as stats per 90, which are the most important.
3. I then analyze the overall game stats for the leagues in question. For club domestic leagues, I use \*just\* that specific league data to predict the number of passes, shots, shots on goal, xg, xgot, and xa based on each team’s rating. For passes, I use the \*quality\*, which is both teams rating combined, to first calculate the number of passes in the game. For Cup games, I use the data for each domestic league that the teams play in as well as Cup data. For international games, I use all international data. Each league that is analyzed must have at least a half season of xg-level statistics in order for the models to be calculated in this way.
4. I do the same thing with the formational data: Based on each team’s formation, I find every game in the past 3 years in which that formation matchup occurs and the team’s have a rating. I then go position row by position row (think defenders, midfielders, attackers) to get a number for every statistic for each position.
5. After that, I get all of the data for the managers -- every game they have managed over the past three years. I create a dataset that has the average number of statistics for each manager \*AND\* their opponents. Based on the average rating difference for each manager, I give them a score for each statistic that is based on how many above or below the expected number they have. For example, if Pep Guardiola’s teams pass 500 times but are, on average, rating 10 points higher than their opponents, and the average team passes 450 times when they have a rating of 10 points or higher, then Guardiola receives a 10/9 boost in the team’s passing statistics.
6. To assess relative home and away strengths, I get the team data for each team and compare their home and away performances to each other’s, adjusting each team’s stats slightly based on whether any team has an above-average advantage either home or away.
7. Finally, I assess the lineups and rate it based on each team’s optimal lineup (based on FotMob rating). Though this is not an exact science, I believe it accounts for players joining clubs who come from incomparable leagues as well as the half-lineups that occur due to injuries or cup-related decisions.

Using all this analysis, I create a dataset that contains each team’s expected number of passes, shots, shots on goal, expected goals, expected goals on target, and expected assists. Each of the ratings from the managers, teams, and lineups gets averaged and then multiplied by the teams’ relative stats to adjust for the factors that go beyond simply league and formational data. I then fit the player stats to the team stats, creating player-level expectations for all the aforementioned stats that are based on a range of factors such as the positional analysis given by the formation data and the relative \*share\* of each statistic that each player has gotten in the past. The last two pieces of score adjustments are the most important, in my opinion, to the outcome of the game that are simply not taken into account in any of the previous analysis: the goalie and the substitutions. So, I rate the goalies and adjust each expected score based on how likely they are to give up a goal based on the opponent’s xg and rating difference. Then, I take away the stats from players who will be subbed off (by the amount of time that they will not be on the field compared to 90 minutes) and add the most likely substitutes based on minutes played in that role (weighted towards recently).

With this dataset of all statistics from starters and substitutions, I create score predictions based on four factors: A linear shot / shot on goal model, a summation of xg, a summation of xgot, and a summation of xa \* the relative levels of xgot / xa. Those four models are then compared to lines I get from the Barstool Sportsbook. Currently, the threshold for a bet is if all models indicate a percentage above the one given by the Barstool Sportsbook; for example, if Lens has a projected a set of (36, 39, 35, 37) percentage chances to win by the models and the price offered is 32, the model suggests a bet on Lens.