Sports Analytics in Football

By - Rohit Singh

Domain of the project

Unveiling the Potential of Sports Analytics: Predicting Football Match Outcomes

In the realm of sports analytics, the meticulous examination of data and application of statistical models are pivotal in unraveling the intricacies underlying football match predictions.

Football, renowned for its worldwide appeal, intricate strategies, and multifaceted dynamics, serves as an ideal domain for exploring the depths of predictive analytics.

Through a fusion of exploratory data analysis and sophisticated statistical modeling techniques, this project endeavors to decode the myriad factors influencing match results, paving the path towards insightful forecasts in the realm of football.

Predictive Question/Hypothesis

Unraveling the Future: Can Team Statistics and Historical Performance Illuminate Football Match Outcomes?

- In our pursuit of forecasting football match results for the upcoming season, we delve into the realm of predictive analytics.
- By leveraging team statistics and historical performance data, our model endeavors to unveil the probabilities associated with various outcomes in a football match.
- At its core, the model harnesses historical goal scoring data as a foundation for projecting the likelihood of diverse match results, offering a glimpse into the potential trajectories of upcoming football encounters.

Target Variables

In our quest to dissect the essence of football match outcomes, we pinpoint two pivotal target variables:

- Goals Per Game Scored:
  - Computed as the ratio of the total number of goals scored to the number of matches played.
  - Serves as a fundamental metric reflecting offensive prowess and attacking efficiency.

- Goals Per Game Conceded:
  - Derived from the division of total goals conceded by the number of matches played.
  - Crucial in gauging defensive solidity and resilience against opposition attacks.

These target variables stand as pillars in our analytical framework, encapsulating the essence of team performance in both offensive and defensive dimensions, thereby laying the groundwork for nuanced predictive insights.

Correlation Variables

In our intricate web of football analytics, we identify a network of correlation variables crucial in illuminating the underlying dynamics of match outcomes. These variables include:

Total Average Home Goals Per Game Scored

Total Average Away Goals Per Game Scored

Average Home Team Goals Per Game Scored at Home

Average Away Team Goals Per Game Conceded Away

Average Away Team Goals Per Game Scored Away

Average Home Team Goals Per Game Conceded at Home

Furthermore, we encapsulate these variables into broader dimensions of team performance, delineating:

Home Attack: Reflecting the offensive prowess of the home team.

Away Defense: Gauging the defensive resilience of the away team.

Away Attack: Measuring the attacking proficiency of the away team.

Home Defense: Assessing the defensive capabilities of the home team.

This array of correlation variables forms the backbone of our analytical framework, elucidating the intricate interplay between offensive and defensive strategies, ultimately contributing to a comprehensive understanding of football match dynamics.

Untangling the DataWeb: Journey in Seeking and Cleansing Data

Data Entry and Web Scraping -

Embarking on our data journey, we meticulously navigate through the vast expanse of football statistics, sourcing our data from esteemed repositories including:

- ESPN Soccer https://www.espn.com/soccer/
- BetExplorer https://www.betexplorer.com/

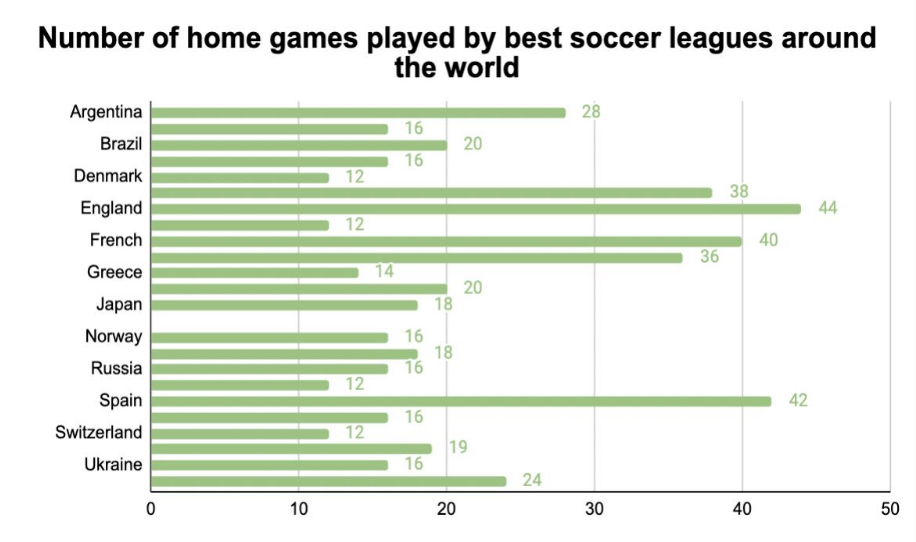
Data Cleaning -

With a keen eye for detail, we embark on the pivotal stage of data cleaning, ensuring the integrity and reliability of our dataset. While our initial data presented a commendable degree of cleanliness, we embarked on several modifications to enhance its utility and clarity:

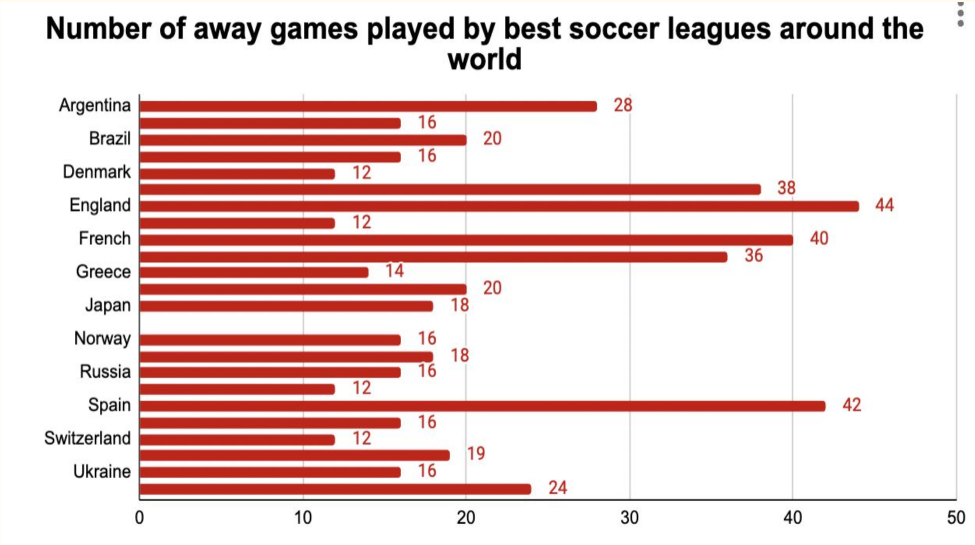
- Restructuring of Column Headings: Tailoring column headings for enhanced comprehension and clarity.
- League Attribution: Augmenting our dataset with a 'League' column, delineating the respective league affiliation of each team.
- Statistical Enrichment: Introducing pivotal metrics such as 'Goals per Game Scored' and 'Goals per Game Conceded', pivotal for Exploratory Data Analysis (EDA) and subsequent modeling endeavors.

Through these meticulous endeavors, we fortify the foundation of our analysis, ensuring that our dataset not only meets but exceeds the stringent standards requisite for insightful sports analytics.

Analysis of Home and Away Games Distribution in Top Soccer Leagues Worldwide



Number of home games played by best soccer leagues around the world



Number of away games played by best soccer leagues around the world

The uniformity in the number of home and away games within each league suggests a prevalent round-robin format. In this format, each team engages in a home-and-away fixture against every other team within the league, ensuring equitable distribution of matches and fostering a balanced competitive landscape. This empirical observation underscores the league organizers' commitment to fair play and integrity, as evidenced by the consistent parity in home and away game allocations across diverse soccer leagues. Such uniformity not only guarantees a level playing field for all participating teams but also enhances the overall excitement and unpredictability inherent in elite soccer competitions.

Navigating the Data Maze with Exploratory Data Analysis

Embarking on our journey through the labyrinth of data, we meticulously organize and dissect the wealth of information collected, structured by team and league affiliations.

At the heart of our analytical framework lies the Poisson distribution, a powerful statistical tool employed to gauge the likelihood of specific events occurring a certain number of times.

In the realm of football analytics, this distribution serves as a cornerstone, facilitating the estimation of goal-scoring probabilities within a match.

Our model harnesses the predictive prowess of the Poisson distribution to forecast the potential outcomes of football encounters.

By calculating the probabilities of various score lines for both the home and away teams, we unravel the intricate tapestry of potential match results, ranging from resounding victories to hard-fought draws.

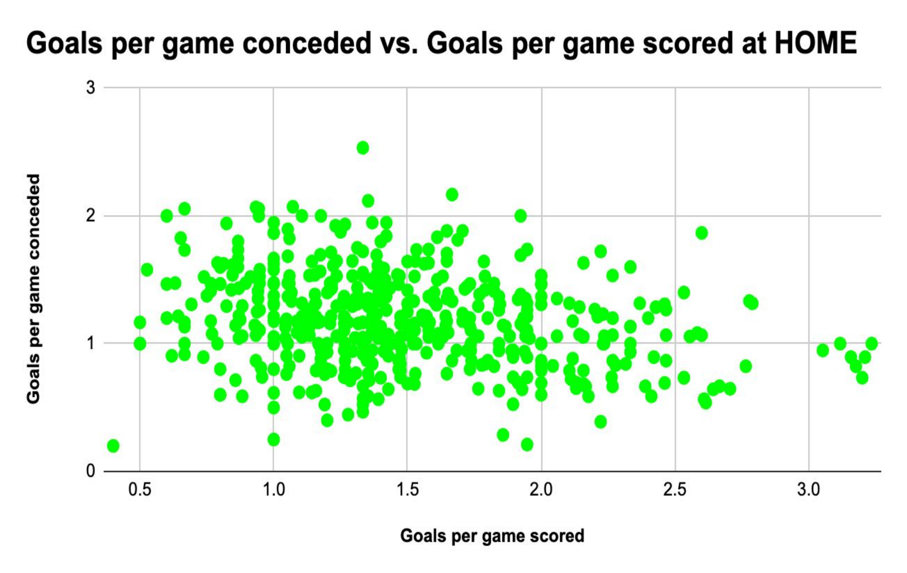
The probabilities of the two teams' scores are then multiplied together to find the probability of each possible outcome (e.g., 0-0, 1-0, 2-2).

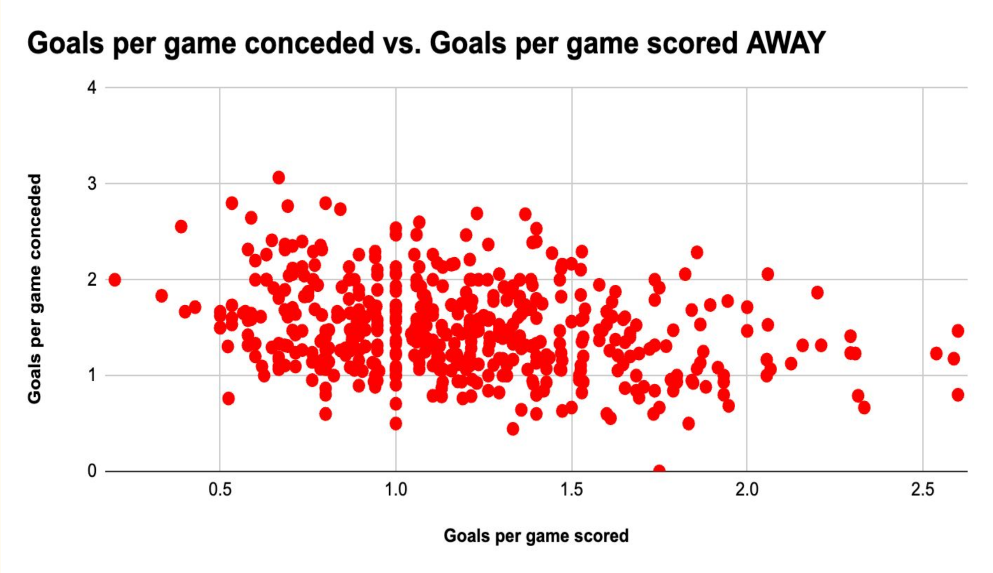
In essence, our model offers a holistic perspective on match projections, encompassing the probabilities of -

- ❖ Home Win
- ❖ Draw
- ❖ Away Win
- ❖ Projected number of Home goals in a match
- ❖ Projected number of Away goals in a match

Through the lens of statistical rigor, we navigate the data maze, illuminating the pathways to informed decision-making and insightful predictions in the realm of football analytics.

Discoveries Post-Analysis: Major Findings

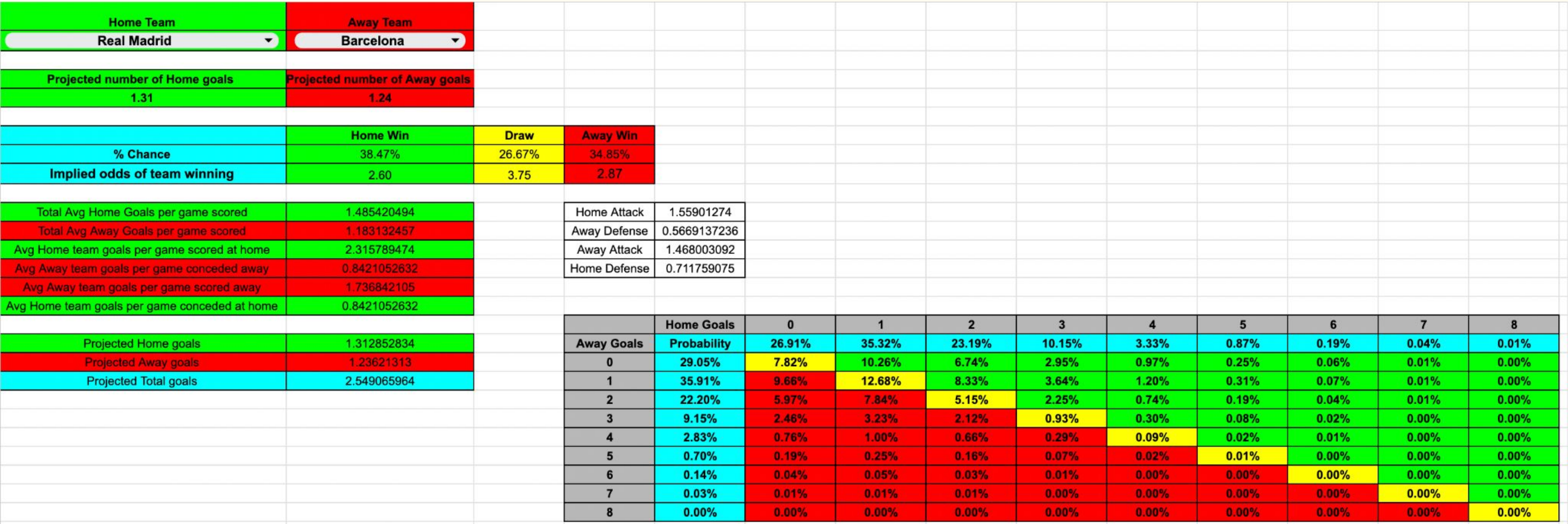




These graphs show two variables: goals per game scored at home and goals per game scored away. Based on the scatter plots, we can derive the following conclusions: Goals per game scored at home appear to be higher than those scored away. This is because the majority of data points in the HOME graph are higher than those in the AWAY graph.There appears to be a positive relationship between goals per game scored at home and abroad. This suggests that a team that scores more goals at home will also score more goals away, and vice versa. However, the association is not particularly strong. Some clubs score more goals at home than abroad, and vice versa.

Scenario: Unveiling Insights from the Clásico Showdown

When Real Madrid Plays at Home



Home Team	Away Team
Real Madrid	Barcelona

Projected number of Home goals	Projected number of Away goals
1.31	1.24

	Home Win	Draw	Away Win
% Chance	38.47%	26.67%	34.85%
Implied odds of team winning	2.60	3.75	2.87

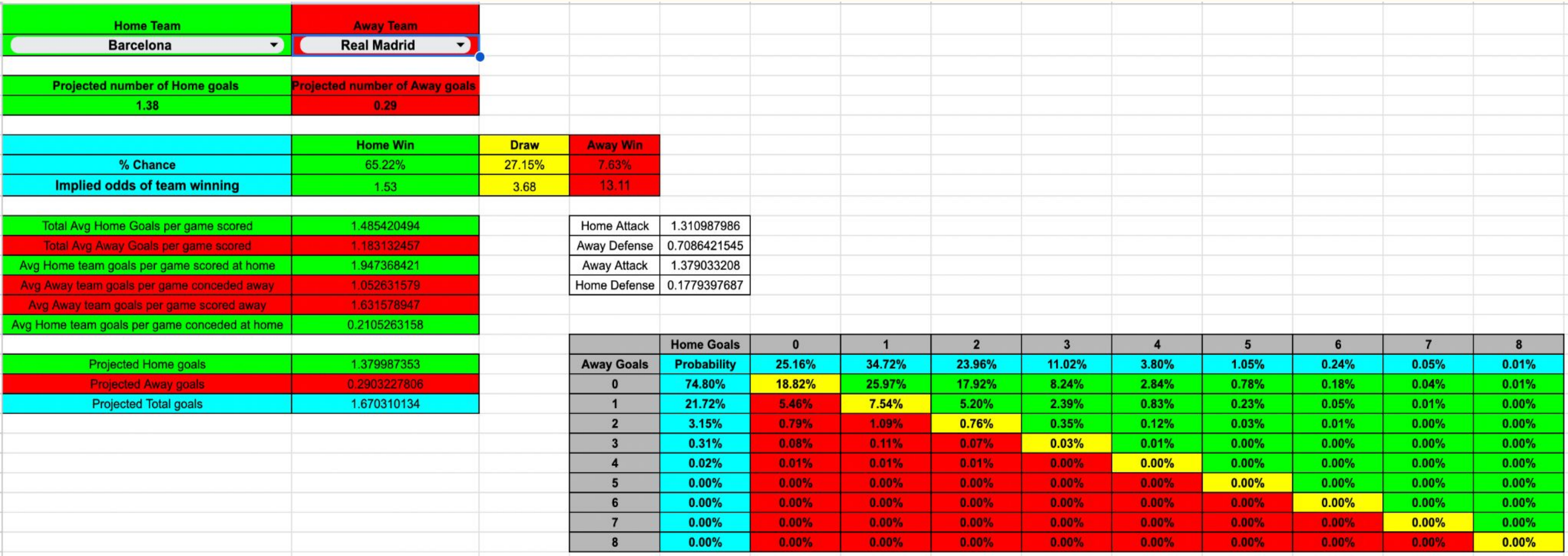
Total Avg Home Goals per game scored	1.485420494
Total Avg Away Goals per game scored	1.183132457
Avg Home team goals per game scored at home	2.315789474
Avg Away team goals per game conceded away	0.8421052632
Avg Away team goals per game scored away	1.736842105
Avg Home team goals per game conceded at home	0.8421052632

Home Attack	1.55901274
Away Defense	0.5669137236
Away Attack	1.468003092
Home Defense	0.711759075

Projected Home goals	1.312852834
Projected Away goals	1.23621313
Projected Total goals	2.549065964

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probability	26.91%	35.32%	23.19%	10.15%	3.33%	0.87%	0.19%	0.04%	0.01%
0	29.05%	7.82%	10.26%	6.74%	2.95%	0.97%	0.25%	0.06%	0.01%	0.00%
1	35.91%	9.66%	12.68%	8.33%	3.64%	1.20%	0.31%	0.07%	0.01%	0.00%
2	22.20%	5.97%	7.84%	5.15%	2.25%	0.74%	0.19%	0.04%	0.01%	0.00%
3	9.15%	2.46%	3.23%	2.12%	0.93%	0.30%	0.08%	0.02%	0.00%	0.00%
4	2.83%	0.76%	1.00%	0.66%	0.29%	0.09%	0.02%	0.01%	0.00%	0.00%
5	0.70%	0.19%	0.25%	0.16%	0.07%	0.02%	0.01%	0.00%	0.00%	0.00%
6	0.14%	0.04%	0.05%	0.03%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.03%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

When Barcelona Plays at Home



Home Team	Away Team
Barcelona ▾	Real Madrid ▾

Projected number of Home goals	Projected number of Away goals
1.38	0.29

	Home Win	Draw	Away Win
% Chance	65.22%	27.15%	7.63%
Implied odds of team winning	1.53	3.68	13.11

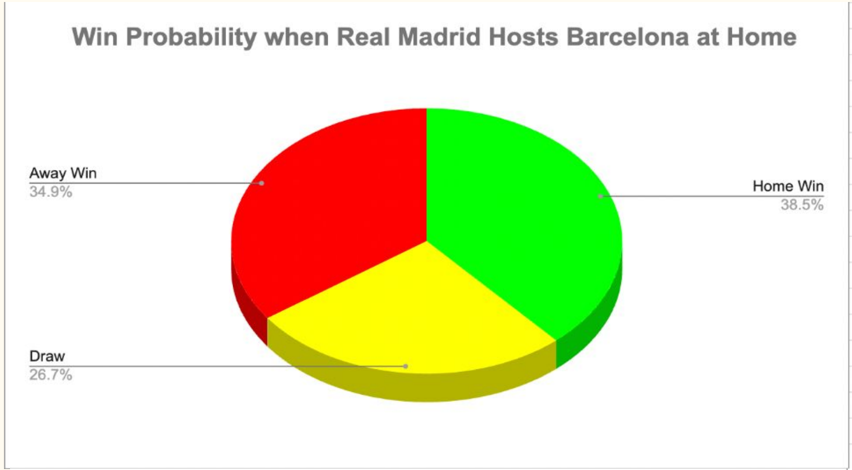
Total Avg Home Goals per game scored	1.485420494
Total Avg Away Goals per game scored	1.183132457
Avg Home team goals per game scored at home	1.947368421
Avg Away team goals per game conceded away	1.052631579
Avg Away team goals per game scored away	1.631578947
Avg Home team goals per game conceded at home	0.2105263158

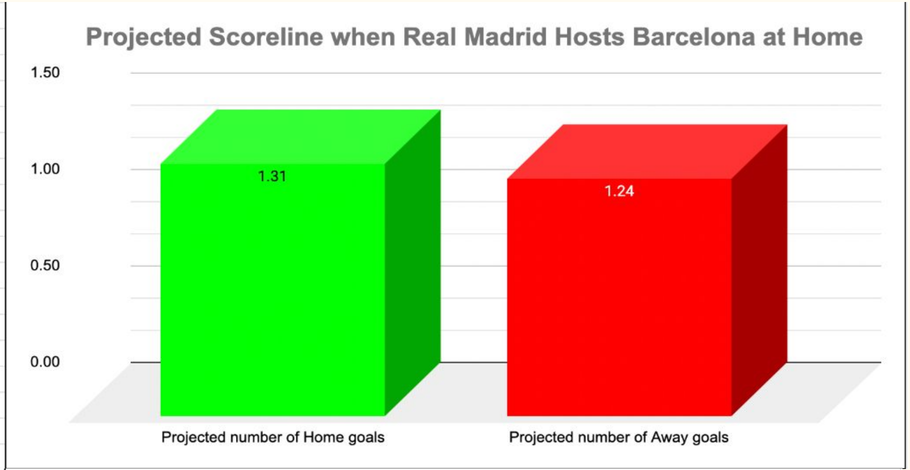
Home Attack	1.310987986
Away Defense	0.7086421545
Away Attack	1.379033208
Home Defense	0.1779397687

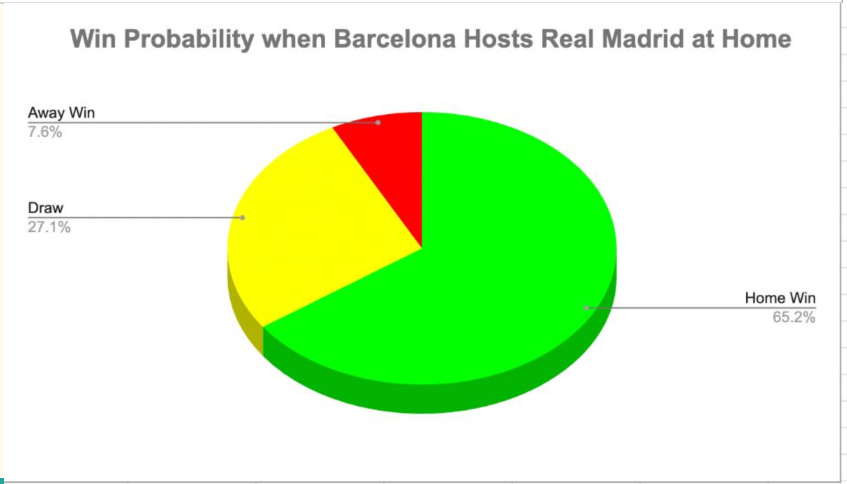
Projected Home goals	1.379987353
Projected Away goals	0.2903227806
Projected Total goals	1.670310134

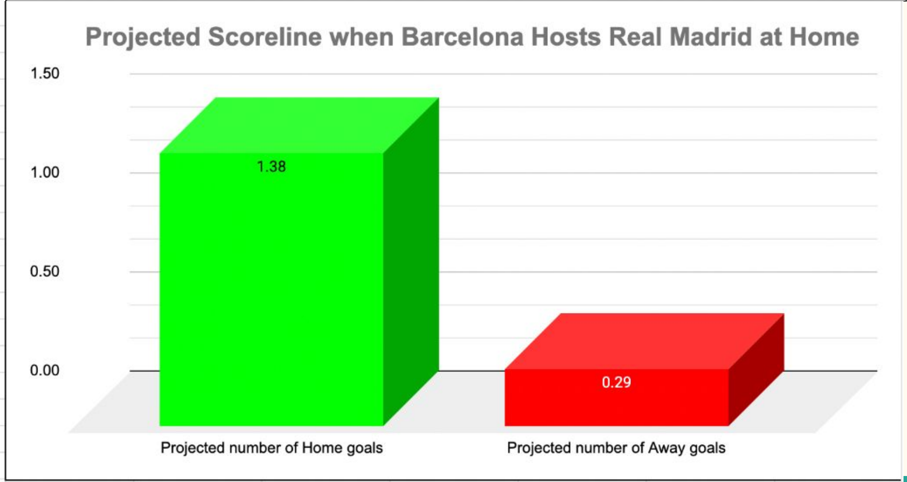
	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probability	25.16%	34.72%	23.96%	11.02%	3.80%	1.05%	0.24%	0.05%	0.01%
0	74.80%	18.82%	25.97%	17.92%	8.24%	2.84%	0.78%	0.18%	0.04%	0.01%
1	21.72%	5.46%	7.54%	5.20%	2.39%	0.83%	0.23%	0.05%	0.01%	0.00%
2	3.15%	0.79%	1.09%	0.76%	0.35%	0.12%	0.03%	0.01%	0.00%	0.00%
3	0.31%	0.08%	0.11%	0.07%	0.03%	0.01%	0.00%	0.00%	0.00%	0.00%
4	0.02%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
6	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Illuminating EDA: Data Visualizations in Focus


Win Probability when Real Madrid Hosts Barcelona at Home


Projected Scoreline when Real Madrid Hosts Barcelona at Home


Win Probability when Barcelona Hosts Real Madrid at Home


Projected Scoreline when Barcelona Hosts Real Madrid at Home

Discoveries Post-Analysis: Major Findings

- The model outputs the predicted probability of each outcome as well as the implied predicted odds for each outcome.
- Following a thorough exploratory data analysis (EDA) combined with extensive visualization resources, a significant finding emerged -
  - Teams fighting on their home grounds have a distinctive and noticeable advantage. The analysis of the data reveals a significant difference in victory probability between home and away games. Particularly, statistical evidence shows that whenever teams play at home, their chances of winning are considerably greater than when they play away.
  - Furthermore, the data reveal unique insights into expected score lines across matchups. Notably, when a team hosts a contest, the expected number of goals, which indicates scoring potential, far exceeds that of their away games. This gap highlights the complicated dynamics that occur in the world of sports competition, as well as the major influence of context-related variables like home-field advantage on match outcomes.

Key learnings

- Improved ability to discern relevant data points from a vast pool of information, ensuring the accuracy and reliability of sports analytics predictions.
- Expanded knowledge of emerging technologies and tools in sports analytics, machine learning frameworks and data visualization software, to stay at the forefront of innovation in the field.
- Increased familiarity with advanced statistical methods like regression analysis & machine learning algorithms, facilitating more precise and nuanced predictions in sports analytics.
- In reference to sports analytics -
  - The older the data used, the less relevant it is for predicting how a team is performing.
  - Using less data makes the model's results less accurate, emphasizing the importance of comprehensive data analysis.