Data and Win Ratio Analysis in European Soccer

University of Colorado Boulder

In today's age, there is a vast amount of data generated from sports, including player skills, game results, seasonal performance, and league management. The challenge for sports science lies in analyzing this data to gain a competitive edge. This data analysis can be approached through various techniques and statistical methods to extrapolate valuable insights. Notably, in recent years, the task of modeling soccer data has gained popularity, with result predictions being a widely explored area. This paper presents a data analysis that predicts the result of future soccer matches based on rank position and historical data. The model was put to test using a data set that comprised the results from 25,000 soccer matches from European Soccer Leagues.

Additional Keywords and Phrases: Bayesian Model, Apriori Algorithm, Cluster analysis, machine learning

1 INTRODUCTION

1.1 Problem

In the realm of sports, especially soccer, large volumes of data are produced. This data encapsulates diverse aspects such as player skills, match outcomes, seasonal games, and league management. A significant challenge that sports science faces today is the pivotal task of analyzing this data in a way that garners a competitive edge. The problem of crafting models for soccer data has been gaining traction in recent years, with result prediction turning into a popular subject.

Knowing that soccer is an intrinsic part of global culture, this predictive model can be an invaluable tool for teams worldwide, eventually driving their strategic and tactical decisions and, subsequently, their winning chances.

2 TASK AND CHALLENGES

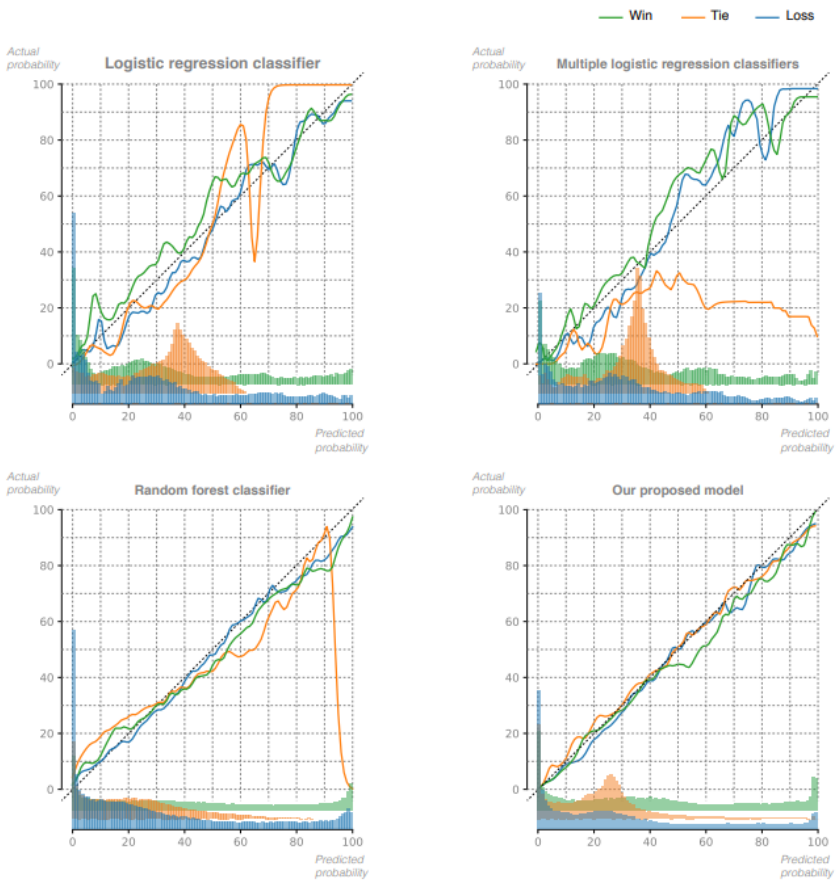
To navigate this challenge, we propose the idea of a model based on the Bayesian theorem. This model considers rank positions and shared history as primary variables for predicting future soccer match outcomes. We have validated the efficacy of this model using a comprehensive data set. This data set consists of the results of 25,000 soccer matches from European Soccer Leagues.

The study intersects match opportunities, player contributions, habitual player combinations, and team attack strategies to carve out insightful deductions. By analyzing Real Madrid's chief attack patterns, for instance, the study crosses the threshold of traditional analysis, offering a robust correlation between the players and their field time. The end goal is achieving a symmetry between data analysis discoveries and real soccer dynamics.

2.1 Review of relevant work

2.1.1 Bayesian Model

Pieter Robbere's study introduce an in-game win probability model for soccer that addresses the shortcomings of existing models. [1]This analysis is performed using various techniques and statistical methods to generate valuable insights. The issue of modeling soccer data has gained popularity in recent years, with predicting game results being the primary focus. This paper introduces a Bayesian Model that predicts future soccer match outcomes based on rank position and historical data. The model was validated using a dataset of over 200,000 soccer matches from leagues worldwide.



2.1 Bayesian model clearly outperforms the LR, mLR and RF models

[1] Pieter Robberechts, Jan Van Haaren, Jesse Davis, 2021. A Bayesian Approach to In-Game Win Probability in Soccer.

2.1.2 Apriori Algorithm

Tao Xie's study focuses on the application of the Apriori algorithm in soccer games, specifically examining data from Real Madrid's five matches during the quarter-finals, semi-finals, and final stages of the 2016-2017 UEFA Champions League season. [2]The assessment focuses on the concept of "scoring opportunities", which are well defined in the literature. The Apriori algorithm is used to explore the relationship between scoring chances, individual players, and frequently observed player groupings. The study discovers a significant bond between the "frequent player combinations" and players responsible for creating scoring opportunities, revealing the optimal player combinations. Moreover, the paper scrutinizes the spatial associations in team offensive strategies, dissecting the primary attacking patterns of Real Madrid. The investigation also delves into the connection between players and their play duration, making sure the data analysis outcomes are consistent with the real-world dynamics of soccer matches.

2.2 Proposed Work

2.2.1 Data and tasks

The ultimate Soccer database for data analysis and machine learning contains a wealth of information, including data from over 25,000 matches and 10,000 players spanning the seasons from 2008 to 2016. It covers 11 European countries and their top championships, providing detailed players and teams' attributes sourced from the FIFA video game series by EA Sports. The dataset also includes team lineups with squad formations represented by X and Y coordinates, betting odds from multiple providers, and comprehensive match events data for over 10,000 matches such as goal types, possession, corners, crosses, fouls, and cards. This dataset is a valuable resource for conducting in-depth analysis and creating machine learning models in the field of soccer.

2.2.2 Tools and techniques

Aim to analyze a comprehensive soccer database encompassing over 25,000 matches and 10,000 players from 11 European countries' top leagues between 2008 and 2016. The dataset includes players' and teams' attributes extracted from EA Sports' FIFA video game series, along with weekly updates. It provides team line-ups with squad formations expressed in X and Y coordinates, betting odds from multiple providers, and detailed match events such as goal types, possession, corners, crosses, fouls, and cards for over 10,000 matches. Our project will involve utilizing this extensive dataset for data analysis and potentially employing machine learning techniques to extract meaningful insights.

In this project, I will utilize the ultimate Soccer database for data analysis and machine learning, which contains detailed information on matches, players, teams, attributes, and match events from the seasons 2008 to 2016 across 11 European countries. I plan to first explore the data to understand player performance, team dynamics, and match outcomes. Next, I will preprocess the data, perform feature engineering, and apply machine learning algorithms to predict match outcomes or player performance. Techniques such as data visualization, statistical analysis, and machine learning models like regression or classification may be employed to derive insights and create predictive models. Python libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib will be used for data manipulation, analysis, and visualization. Ultimately, the goal is to extract valuable insights from the dataset and build predictive models that can enhance decision-making in the world of soccer.

[2] Tao Xie, Yaxian Hao, Yanyuan Xing. 2023. Application of the Apriori Algorithm in Soccer Games.

(1) Cumulative historical average: This approach involves calculating the average performance of a team up to a certain point in a game based on historical data.

(2) Logistic regression classifier (LR): This model uses a logistic regression algorithm to predict the probability of a specific outcome (win, tie, loss) for the home team based on game state features and historical match data.

(3) Multiple logistic regression classifiers (mLR): This model divides game state data into separate time frames and trains different logistic regression classifiers for each time frame to account for the impact of time remaining on the win probability.

(4) Random forest model (RF): This model uses a random forest algorithm to capture non-linear interactions between all game state variables, providing a more comprehensive analysis of the data.

Many win probability models in sports games focus on predicting the likelihood of the home team winning directly through machine learning. However, our approach involves modeling the number of goals a team will score in the future and then using that information to determine the win-draw-loss probability. This means we predict the probability distribution of goals scored by each team from the current game state until the end of the match.

2.3 Evaluation

For the project to measure the results and define success, the outlined evaluation plan should be observed. Success in this project would involve a high prediction accuracy, significant improvement over time, and better or similar performance as compared to other models. Furthermore, the model should seamlessly integrate with the current systems and assist in decision-making processes.

2.3.1Model Validation

Once the model has been trained with a subset of our data, we will test it against a set of data that it hasn't seen during its training phase. This will give us insights into the effectiveness and accuracy of our model..

2.3.2 Prediction Accuracy

By comparing our model's predictions with actual outcomes, we can evaluate its performance. This will be measured using the win ratio, which indicates the percentage of matches where the predicted outcome was the actual outcome..

2.3.3 Improvements Over Time

The model must show progressive improvement in prediction over time. This is to ensure that it learns from its past predictions.

2.3.4 Comparison With Other Models

To assess its effectiveness, the model's performance can be compared with other similar models or standard statistical analysis.

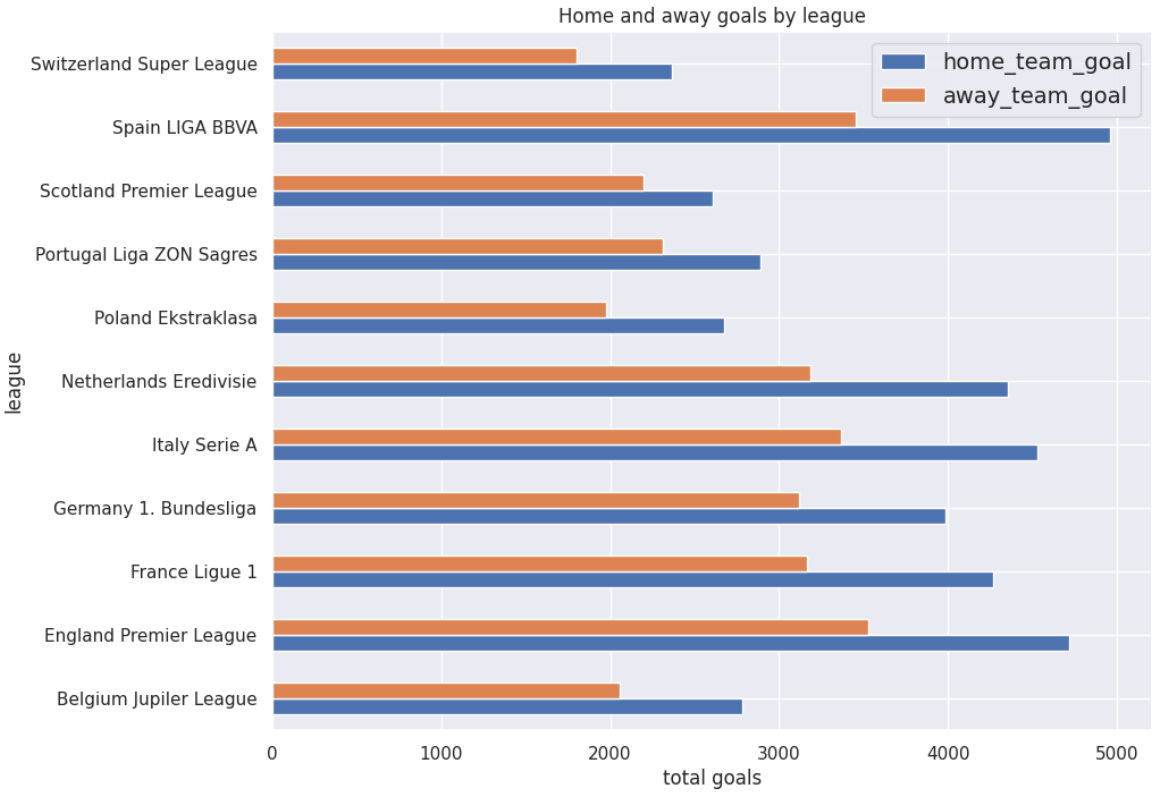
2.4 Timeline

Based on the project timeline, I have completed the Project Proposal in Weeks 2. I am currently working on the Project Checkpoint. Following the completion of the report, I will focus on creating slides for the presentation. I plan to submit

both the report and slides this week, but I will not simultaneously work on them. Once both submissions are completed and I have reviewed my peers' work, I will move on to Week 5 where I will prepare a video presentation to showcase my final slides to my classmates.

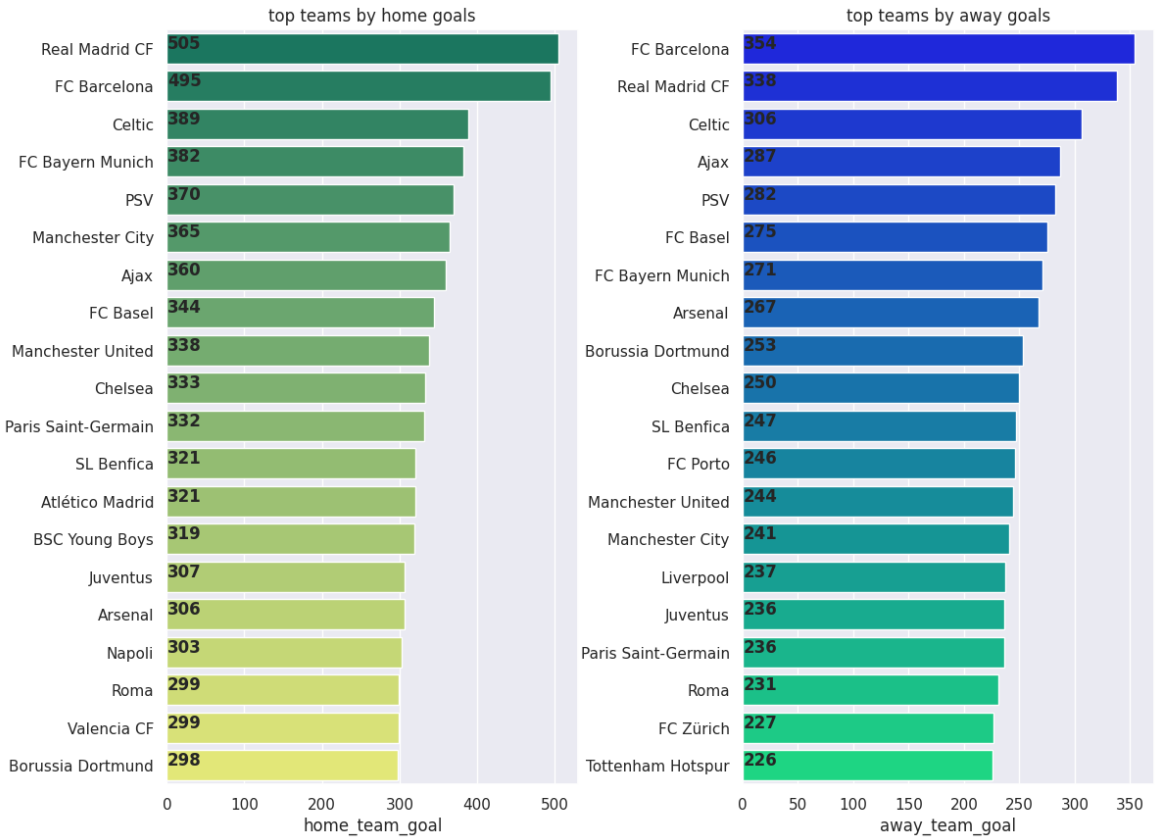
3 TASK AND CHALLENGES

3.1 Exploratory data analysis



3.1 Home and away goals by league

In every league, the home or hosting team is considered to have a significant advantage over the away or visiting team.



3.2 Home and away top teams by goals

3.3 Home and away goals correlation heatmap

3.2 Bayesian data analysis

Many win probability models in sports games focus on predicting the likelihood of the home team winning directly through machine learning. However, our approach involves modeling the number of goals a team will score in the future and then using that information to determine the win-draw-loss probability. This means we predict the probability distribution of goals scored by each team from the current game state until the end of the match.

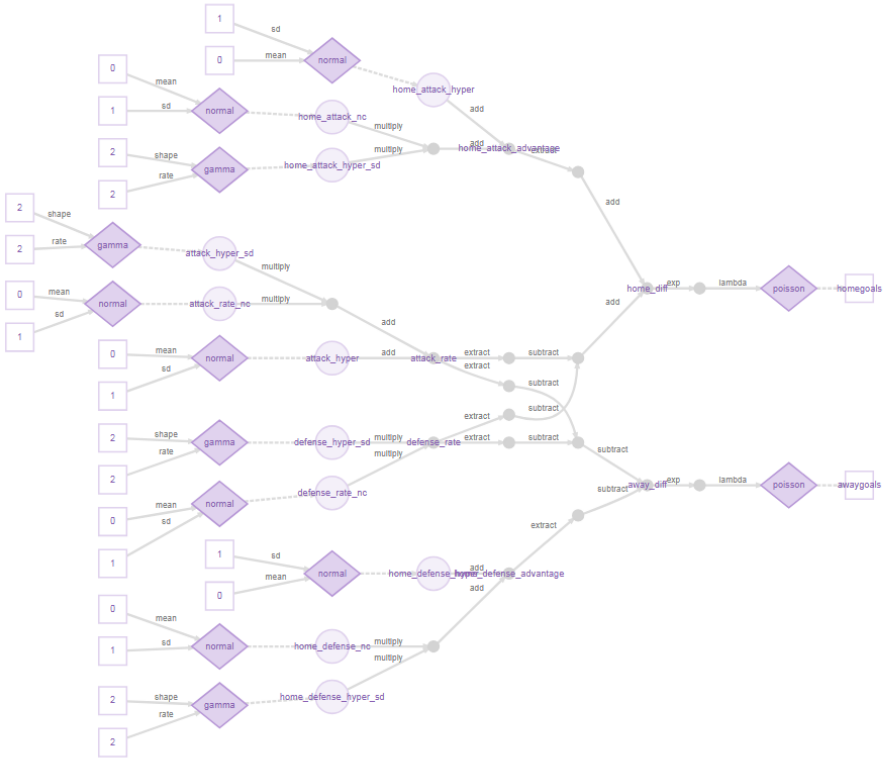
We are using a Bayesian multilevel model in pymc3 to predict the number of goals each team will score in a football match. This model treats goals scored by each team as independent Poisson processes, considering factors like attacking and defense rates along with home advantage. By incorporating all these components in a probabilistic framework, we aim to accurately estimate the distribution of future goals and ultimately the win-draw-loss probabilities in a match:

$$homegoals \sim Poisson(\exp(homediff))$$
$$awaygoals \sim Poisson(\exp(awaydiff))$$

The scoring intensities in a soccer game are estimated based on the game state features of the home and away teams. These features have different levels of importance at different points in the game. In the beginning, the prior strengths of the teams influence the predictions more, whereas towards the end of the game, the features related to the in-game performance become more significant. The variation in importance of these features is not constant and can be influenced by factors like a fierce final sprint in the game. To capture this dynamic process, we model the scoring intensity parameters as a temporal stochastic process. This approach allows us to share information across different time frames and make accurate predictions even for rare events, such as a red card being given in the first minute of the game. Likelihood:

$$\begin{aligned}
homediff &= AttRate_{home} - DefRate_{away} + HomeAttAdvantage_{home} \\
awaydiff &= AttRate_{away} - DefRate_{home} - HomeDefAdvantage_{home} \\
AttRate_i &= BaseAtt + AttRate_{i,non-centered} * \tau_i^{att} \\
DefRate_i &= DefRate_{i,non-centered} * \tau_i^{def} \\
HomeDefAdvantage_{home} &= BaseHomeAtt + HomeAttRate_{i,non-centered} * \tau_i^{HomeAtt} \\
HomeDefAdvantage_{home} &= BaseHomeDef + HomeDefRate_{i,non-centered} * \tau_i^{HomeDef} \\
BaseAtt &\sim Normal(0,1) \\
AttRate_i &\sim Normal(0,1) \\
DefRate_i &\sim Normal(0,1) \\
BaseHomeAtt &\sim Normal(0,1) \\
HomeAttRate_i &\sim Normal(0,1) \\
BaseHomeDef &\sim Normal(0,1) \\
HomeDefRate_i &\sim Normal(0,1) \\
\tau_i^{att} &\sim Gamma(2,2) \\
\tau_i^{def} &\sim Gamma(2,2) \\
\tau_i^{HomeAtt} &\sim Gamma(2,2) \\
\tau_i^{HomeDef} &\sim Gamma(2,2)
\end{aligned}$$

I have developed a Bayesian model that enables the analysis of different factors such as attack hyper, attack hyper standard deviation, attack rate, defense rate non-championship, defense hyper standard deviation, defense rate, home attack hyper, home attack hyper standard deviation, home attack non-championship, home attack advantage, home defense hyper, home defense hyper standard deviation, home defense non-championship, and home defense advantage:



3.4 Bayesian model

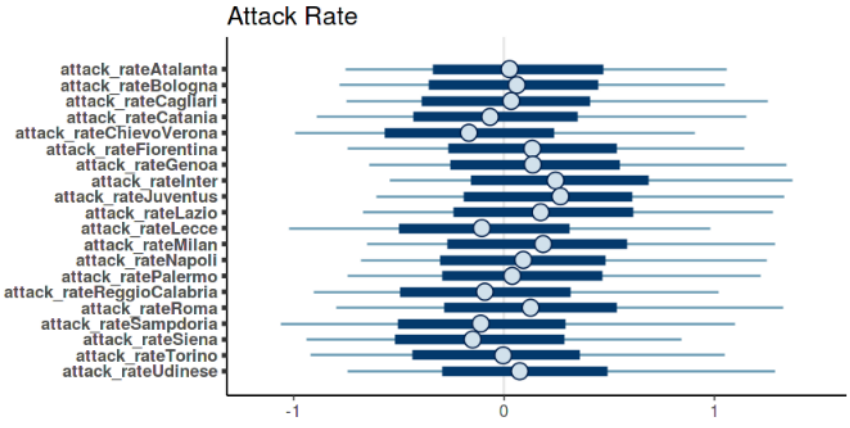
Parameters of the Bayesian model:

	parameter	y	ll	l	m	h	hh
attack_rateAtalanta	attack_rateAtalanta	20	-0.7518963	-0.3383055	0.026833002	0.4721370	1.0597909
attack_rateBologna	attack_rateBologna	19	-0.7825846	-0.3563162	0.060745187	0.4462552	1.0485524
attack_rateCagliari	attack_rateCagliari	18	-0.7494888	-0.3917186	0.034217771	0.4103758	1.2521037
attack_rateCatania	attack_rateCatania	17	-0.8874943	-0.4288318	-0.066187948	0.3519075	1.1508319
attack_rateChievoVerona	attack_rateChievoVerona	16	-0.9937813	-0.5677773	-0.166654222	0.2375559	0.9068941
attack_rateFiorentina	attack_rateFiorentina	15	-0.7408080	-0.2622603	0.134495676	0.5345383	1.1405900
attack_rateGenoa	attack_rateGenoa	14	-0.6396995	-0.2530506	0.137036227	0.5520175	1.3422012
attack_rateInter	attack_rateInter	13	-0.5444802	-0.1551161	0.245253474	0.6900044	1.3724558
attack_rateJuventus	attack_rateJuventus	12	-0.6040236	-0.1923763	0.268472224	0.6096458	1.3321803
attack_rateLazio	attack_rateLazio	11	-0.6711828	-0.2394964	0.174283907	0.6158894	1.2766708
attack_rateLecce	attack_rateLecce	10	-1.0194794	-0.4980956	-0.105199948	0.3101372	0.9821960
attack_rateMilan	attack_rateMilan	9	-0.6526977	-0.2693751	0.186235264	0.5848321	1.2860808
attack_rateNapoli	attack_rateNapoli	8	-0.6807079	-0.3025428	0.091812849	0.4826207	1.2503573
attack_ratePalermo	attack_ratePalermo	7	-0.7437713	-0.2926718	0.039273847	0.4684222	1.2198950
attack_rateReggioCalabria	attack_rateReggioCalabria	6	-0.9020205	-0.4921896	-0.091105454	0.3174626	1.0204052
attack_rateRoma	attack_rateRoma	5	-0.7991127	-0.2855074	0.126511455	0.5370904	1.3278993
attack_rateSampdoria	attack_rateSampdoria	4	-1.0616322	-0.5020528	-0.110529732	0.2929896	1.0978253
attack_rateSiena	attack_rateSiena	3	-0.9388392	-0.5205768	-0.148960054	0.2864299	0.8417343
attack_rateTorino	attack_rateTorino	2	-0.9169977	-0.4340753	-0.003741728	0.3597125	1.0495547
attack_rateUdinese	attack_rateUdinese	1	-0.7448290	-0.2960373	0.074977241	0.4904186	1.2872247

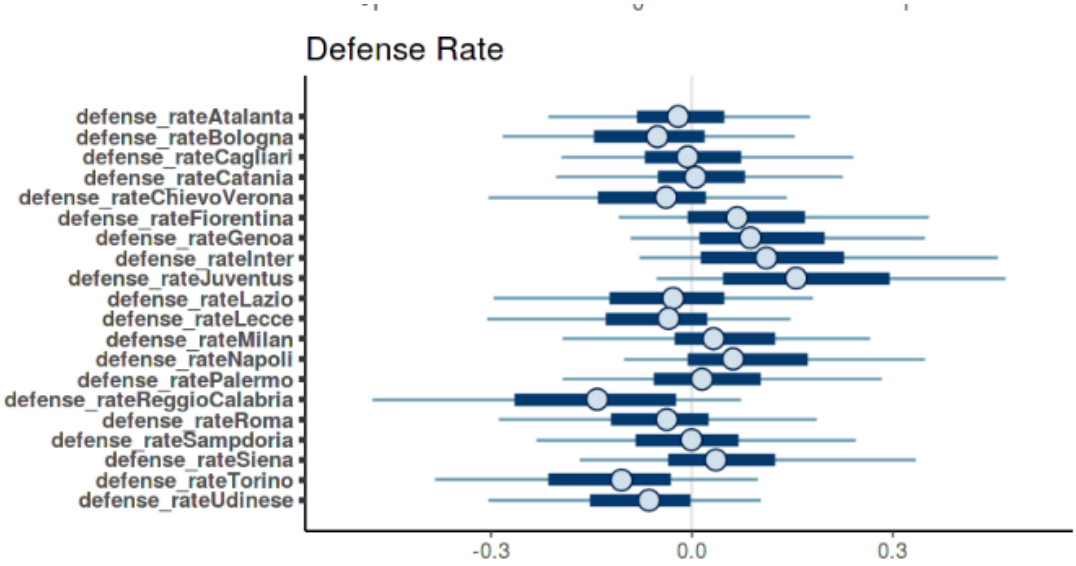
3.5 Bayesian model parameters

3.3 Evaluation and feature importance

The Bayesian model is proficient in consistently making accurate predictions by considering various game features throughout each match. Initially, its performance resembles that of a basic logistic regression model using Elo rating as a single feature. However, the Bayesian model surpasses LR, mLR, and RF models in terms of predictive accuracy.



3.6 Bayesian model attack rate evaluation



3.7 Bayesian model defence rate evaluation

When analyzing the impact of different features on win probability estimates, the Bayesian model benefits from the inclusion of base features (game time and score differential), team strength features, and contextual features. These feature groups collectively enhance the Relative Probability Score (RPS), with pregame strength features enhancing early game prediction accuracy. As the game progresses, the addition of contextual features significantly improves prediction accuracy. Towards the end of the game, the influence of team strength and contextual features diminishes, and the current score differential becomes a reliable predictor of the final outcome.

5 CONCLUSIONS

The article discussed the implementation of a Bayesian model for predicting in-game win probabilities in soccer matches. This model utilizes team-specific features and predicts the future number of goals scored by each team as a temporal stochastic process. One notable aspect of this model is that it requires only a relatively small amount of data to learn effectively. In the era of Big Data, where not all datasets are large, this technique offers a reliable way to create robust models while providing insights into their uncertainty. The interpretability of the model parameters and the functionality provided by tools like the NUTS sampler and pymc3 make Bayesian multilevel modeling a powerful and valuable tool for sports analytics.

REFERENCES

[1] Pieter Robberechts, Jan Van Haaren, Jesse Davis, 2021. A Bayesian Approach to In-Game Win Probability in Soccer , KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data MiningAugust , 2021, Pages 3512–3521,

https://doi.org/10.1145/3447548.3467194.

[2] Tao Xie, Yaxian Hao, Yanyuan Xing. 2023. Application of the Apriori Algorithm in Soccer Games . AAIA '23: Proceedings of the 2023 International Conference on Advances in Artificial Intelligence and ApplicationsNovember 2023, Pages 54–59, https://doi.org/10.1145/3603273.3631190.