Soccer Prophet – Machine Learning Techniques for Accurate Match Predictions

Introduction

Soccer Prophet is a machine-learning model that predicts the outcome of soccer matches based on historical data. The model uses various features that occur in the game, such as shots, corner kicks, and fouls to predict whether the match will end in a 'Home Win', 'Away Win', or 'Draw'. Soccer Prophet is trained on a large dataset and was fitted using Logistic Regression, SVM, and XGBoost to determine the best results.

Problem Statement

The underlying question behind the project is: How might we identify metrics that predict soccer match results in order to provide insights that help individuals and organizations make better decisions related to betting and team selection? In the real world, bookies can use the model's predictions to adjust odds and maximize profits, clubs can use the predictions to select lineups, make tactical adjustments and improve performance, and fans can benefit by making more informed gambling decision when placing bets to enhance their enjoyment.

Background on Data use in Soccer

Over the last decade, soccer has become more data oriented. There has been investment into data analysis, with data scientists and analysts working directly with coaches to help them make sense of collected data. In 2015, EPL side Liverpool FC hired a team of data scientists to help them analyze player performance.[1] There is an adoption of data driven decision-making in soccer clubs, where data is analyzed to find areas of strengths and weaknesses with solutions implemented to address any issues. Manchester City, another EPL team, uses data analytics to identify weaknesses in opposing teams and developing tactics to gain an edge in their match.[2] With the rise of analytics, companies such as Opta and Statsbomb provide detailed analysis services to soccer teams, helping them gain a competitive edge through data.[3] It's no question that the need for Data Science in soccer is high in demand as teams look to gain an edge by investing in data acquisition and hiring talent to make sense of it.

[1] "How data analytics is helping Liverpool FC get ahead in the Premier League" (ComputerWeekly, 2017) -

[2] "How Manchester City are using data to gain an edge on their Premier League rivals" (Sky Sports, 2020) - https://www.skysports.com/football/news/11679/12137554/how-manchester-city-are-using-data-to-gain-an-edge-on-their-premier-league-rivals

[3] "Why the future of football is all about data analysis" (The Guardian, 2018) - https://www.theguardian.com/football/2018/jan/25/why-the-future-of-football-is-all-about-data-analysis

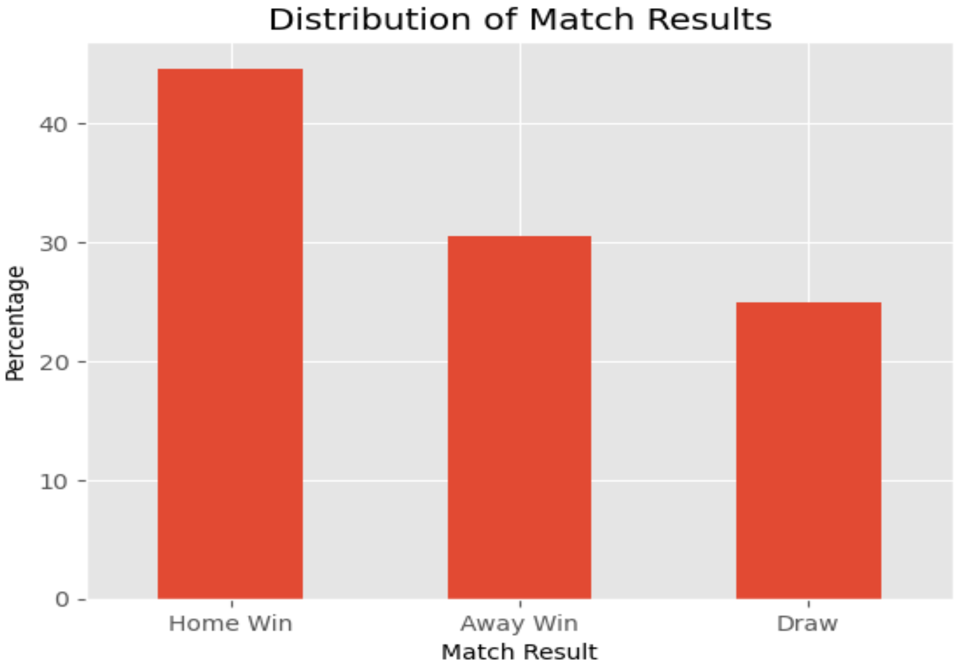
The Dataset

The dataset "Euro Football Data – 2012 to 2023", is a "self-updating" dataset with matches added periodically, is compiled by Georgi Shopov. It originally contained data from 17 leagues in 8 European countries, with 5 of the leagues in second division. The final modeling dataset was reduced to the 5 most viewed leagues in the world, English Premier League, German Bundesliga, Spanish La Liga, French Ligue 1, and Italian Serie A. Shopov compiled this data by scraping Football-Data.co.uk, a free betting, scores, and results service.

Cleaning and Processing

The dataset had a lot of null values in key categories including shots, corner kicks, yellow cards that were prevalent for teams in second divisions. The nulls were also all predominant in the 2012-2013 and 2022-2023 season. The solution was to drop unpopular leagues and the two seasons to reduce the nulls. After that, we were left with a handful of rows with nulls that were manually filled in by searching Football-Data website, thus the uniformity of the dataset was preserved in terms of each seasons completeness. The target variable was "Results" and contained 3 classes: 'Home Win', 'Away Win', and 'Draw'. The results are summarized in Figure 1, with Home Win having 45% of the results, Away Win having 30% and 25% of matches ending in a Draw.

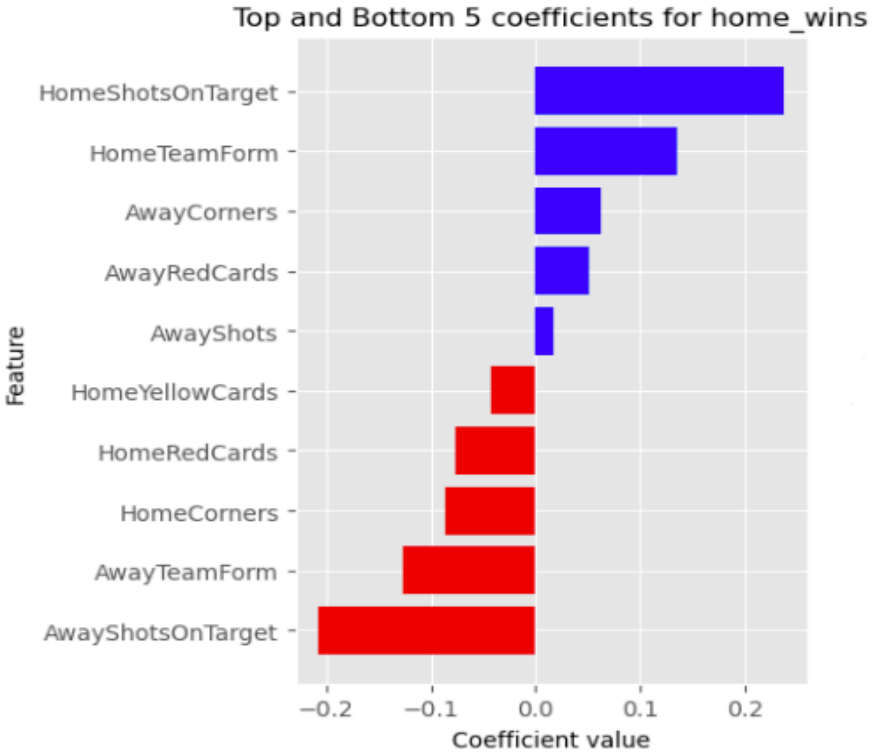
Figure 1: Distribution of Match Results: There is a clear advantage when playing at home.



Exploratory Data Analysis was performed on the teams and split up by league to help identify patterns. An interesting observation was that Barcelona was the most successful club in the dataset with most goals scored when playing at home and away with 496 and 370 respectively. Soccer fans know that this is due to Barcelona's golden generation with prime Lionel Messi, who helped Barcelona win 5 Spanish titles from 2012 – 2019[4]. An interesting finding was that their rival, Real Madrid, was 5th in goals scored when playing at Home but second when playing Away. They also recorded more Away wins than Barcelona in the dataset, an impressive feat considering the disadvantage of playing Away.

At initial modeling, draws were not being predicted or recalled often so a "DrawsInLast3" feature was made to assist the models predictions. Unfortunately, the model began over-predicting draws and reducing the overall accuracy, so the feature was scrapped. One feature that helped improve accuracy was the "Form" category, which calculated the amount of goals scored and subtracted them from the goals conceded in the last 5 games. When looking at the coefficients, it was the second highest feature that helped predict Home and Away wins. Another idea was to split the Season into 3 parts: beginning, middle, and end, but it unfortunately did not improve accuracy.

Figure 2: Top/Bottom 5 coefficients for HomeWins. "HomeTeamForm" is second, behind "HomeShotsOnTarget".



In Figure 2, we can see the predictive coefficients that were most and least predictive for the Logistic Regression model. Interestingly enough, the away team having more corner kicks was more predictive for home wins. This may be due to a frustrated opponent having their shots blocked and deflected by the keeper as they struggled to score. Sometimes it just doesn't want to go in.

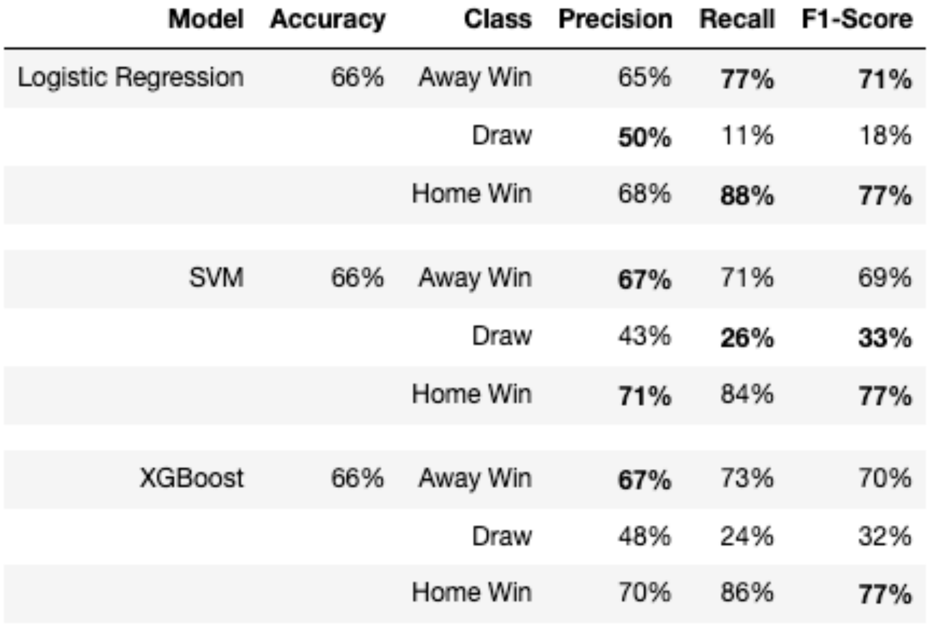
[4] https://www.wikiwand.com/en/List_of_Spanish_football_champions

Insights, Modeling, and Results

Three models were used to make predictions on the dataset: Logistic Regression, SVM, and XGBoost. The accuracy for all 3 models on the test set was 66%, which is good considering that Home Win's represent 45% of the Results. Logistic Regression did well on recalling home and away wins but struggled with draws. Perhaps this is because it's more suited for predicting binary values. Support Vector Machines (SVM), a distance based model, did much better in predicting Draws than Logistic Regression. Though SVM predicted less Home and Away wins overall, when it made the guess, it was correct more often. Logistic Regression had more false positives and false negatives in the win classes than SVM, mostly because it didn't predict many draws. XGBoost performed closely to SVM as it achieved similar scores on predicting wins, having less false negatives in those classes. It predicted more wins than SVM but also struggled with draws.

Figure 3 contains a summary of the 3 models, showing that Logistic Regression performed well on predicting Away Wins, while SVM and XGBoost fared better with Draws and Home Wins. This is seen in the F1 score, a harmonic mean between precision and recall, with recall being percentage of correct predictions of the class and precision defined by how often the model was right when it predicted a class. Logistic Regression has a high recall in the wins column but made a lot of false positives along the way.

Figure 3: All models had 66% accuracy though Logistic Regression did well with predicting wins while SVM and XGboost were better at draws.



Model	Accuracy	Class	Precision	Recall	F1-Score
Logistic Regression	66%	Away Win	65%	**77%**	**71%**
		Draw	**50%**	11%	18%
		Home Win	68%	**88%**	**77%**
SVM	66%	Away Win	**67%**	71%	69%
		Draw	43%	**26%**	**33%**
		Home Win	**71%**	84%	**77%**
XGBoost	66%	Away Win	**67%**	73%	70%
		Draw	48%	24%	32%
		Home Win	70%	86%	**77%**

Findings and Conclusions

It’s not surprising that Draws are difficult to predict since they mostly come down to luck. Though a team can win comfortably by 2 or 3 goals, a draw can occur when a team leading by 1 concedes in the last minute of a game. Perhaps a feature that combines the time of goal with the score at that moment can be helpful in predicting them. Overall, the models did fairly well in predicting Wins but the data set can be expanded to include more events, such as posts hit, time of goal, suspensions, and team line ups. In the future, the dataset will continue being updated and new categories will be scraped from Football-Data to optimize the accuracy of the model. Gambling odds will be added as it’s important to know the betting predictions and what the model can learn from them. As of now, Soccer Prophet is not yet ready to assist bookies and soccer clubs in optimizing their business, but with more matches and features added, it will continue improving in accuracy until it reaches its full potential.