Diep Vu

Yousef Lari

CSC542

Predictive Modeling for Sports Betting: Soccer Case

1. **Introduction:**

Our project is situated in the rapidly growing field of sports analytics, specifically focusing on predictive models for sports betting. In many countries, bookmaking (the profession of accepting sports wagers) is regulated but not criminalized. In the United States, it was previously illegal under the Professional and Amateur Sports Protection Act of 1992 (PASPA) for states to authorize legal sports betting, hence making it effectively illegal. However, the Act was later struck down by the Supreme Court later in 2018. As of 2024, sportsbooks are legal in 37 states in the District of Columbia, while online sports betting is also legal in 30 states and Washington D.C. According to the American Gaming Association's (AGA) Commercial Gaming Revenue Tracker, sports betting achieved new records for handle ($119.8B) and sportsbook revenue ($10.9B), up 27.8 and 44.5 percent respectively ("2023 Commercial Gaming Revenue Reaches $66.5B, Marking Third-Straight Year of Record Revenue"). This multi-billion dollar industry also contributes billions of additional tax dollars to states each year in the form of income, sales, payroll, and various other corporate taxes. Moreover, around one in five U.S. adults (19%) say they have personally bet money on sports in some way in the last 12 months, whether with friends or family, in person at a casino or other gambling venue, or online with a betting app, according to a new Pew Research Center survey (Gramlich). As sports betting becomes legal in additional states, more Americans will gain the opportunity to participate – with tens of millions of people placing bets on everything from college sports to predicting NFL quarterback performance. As a result, there is a burgeoning interest in leveraging statistical analysis to inform betting strategies.

In this research, we concentrate on predicting soccer match outcomes and evaluating relationships between performance metrics and game results by utilizing a comprehensive dataset from Football-Data. Of the total amount bet at sportsbooks, 65-80% is being wagered on soccer due to the sport's immense popularity in the sports betting industry (Lindner). During the 2020-2021 season, bets on Premier League matches alone generated over 68.5 billion euros worldwide. Meanwhile, the global amount wagered on La Liga reached roughly 42.1 billion euros ("Total amount wagered on European soccer worldwide in the 2020/2021 season, by league"). Therefore, it becomes a tantalizing prospect for investors and stakeholders, shining a spotlight on its potential within the sports betting market. This field of predictive models in sports betting holds significant importance due to its potential to revolutionize betting strategies through data-driven insights. By accurately predicting sports outcomes, stakeholders can make more informed decisions, thereby elevating the sports betting industry to new levels of sophistication and success. Our objectives include developing a predictive model that integrates a variety of data points to forecast match outcomes, identifying the most significant factors determining match results, and comparing the effectiveness of different statistical and machine learning methods in sports outcome predictions.

In the remainder of this report, we first discuss the state of the art in Section 2 and describe our data and variables as well as summarize the methods in Section 3 and analyze the results in Section 4. Section 5 contains the conclusions and potential ideas for future work.

2. **State-of-the-art**:

Several studies have explored various approaches and methodologies to develop accurate predictive models. Despite the seemingly limited field of sports, the subject of predicting the outcome of sports events does not only vary in terms of the sport: American football (Song et al), basketball (Loeffelholz et al), European football (Sibony et al), or tennis (Wilkens); but also the competition: national leagues (Angelini and de Angelis), international

tournaments, or Olympics. Recent attempts to solve similar problems involve advancements in neural networks such as feed-forward, radial basis, probabilistic, and generalized regression neural networks (Loeffelholz et al). Fusion of the neural networks is also examined using Bayes belief networks and probabilistic neural network fusion (Loeffelholz et al). Their best model achieved a remarkably high accuracy of over 74% using neural network models, however, their dataset consisted of only 620 games. Another attempt was made by Ivanković, Racković, Markoski, Radosav, and Ivković by using Artificial Neural Networks (ANNs) to predict outcomes of basketball games with high accuracy by formalizing the effects of shots from different court areas as features; however, their specific dataset (League of Serbia in seasons 2005/06 – 2009/10) makes it impossible to compare the results with other research.

Researchers have also explored techniques like attention mechanisms to better capture relevant information from historical football data. Additionally, there is a growing focus on integrating external data sources, such as player transfer news, social media sentiment analysis, and betting market data, to enhance predictive models. Sinha, Dyer, Gimpel, and Smith made use of Twitter posts to predict the outcomes of NFL games. Information from Twitter posts enhanced forecasting accuracy, moreover, a model based solely on features extracted from tweets outperformed models based on traditional statistics. Another research used play-by-play data to develop new features. The main reason why features derived from such data are superior to box score statistics is that they include a context. Out of Naive Bayes, Logistic Regression, Bayes Net, SVM, and KNN, the SVM performed best, achieving an accuracy of over 71% in the course of 10 NBA seasons from 2003/04 to 2012/13 (Puranmalka).

3. **Materials and methods**:

3.1 Dataset:

The data was collected from the website Football-Data.co.ok. Each league season was collected separately by downloading the CSV files and loaded into a data frame using Pandas in Python. There are 43 files from different leagues in 5 countries (England, Germany, Italy, Spain, and Portugal) in 4 seasons from 2020 to 2024, including the current seasons. The data did not include the league place, so we had to generate the points and rank for each team before each match based on the number of wins, draws, and losses as well as the goal difference of each team. Initially, there were 20921 matches with 111 variables. However, we decided to only keep 38 columns with the necessary information. Regarding the betting odds, we only keep the average odds instead of individuals from different bookmakers. After dropping those columns, there were 20 matches with null values, so we dropped all of them since the dataset has more than 20000 values, which is enough for the scope of this project. It is important to note that the odds are in decimal format representing the amount a winning bet would collect on a $1 bet. After performing correlation analysis on the remaining 38 variables, average odds and closing odds are highly correlated, so closing odds were removed. We decided to keep only pre-match average odds and pre-match information, which resulted in the final dataset of 20,901 matches with 21 variables.

Table 1. Variable Description

Variable	Description
Div	Categorical, league division
HomeTeam	Categorical, home team
AwayTeam	Categorical, away team
FTR	Numerical, full-time result (0=Home Win, 1=Draw, 2=Away Win)
HS	Numerical, home team shots
AS	Numerical, away team shots
HST	Numerical, home team shots on target
AST	Numerical, away team shots on target
AvgH	Numerical, market average home win odds
AvgD	Numerical, market average draw odds
AvgA	Numerical, market average away win odds
Avg>2.5	Numerical, market average over 2.5 goals
Avg<2.5	Numerical, market average under 2.5 goals
Ahh	Numerical, market size of handicap (home team)
AvgAHH	Numerical, market average Asian handicap home team odds
AvgAHA	Numerical, market average Asian handicap away team odds

Summary Statistics of Categorical Variables

1. Div - 11 unique values

- England: Premier League (E0), Championship (E1), League 1 (E2), and League 2 (E3)
- Spain: La Liga Primera (SP1), La Liga Segunda (SP2)
- Germany: Bundesliga 1 (D1), Bundesliga 2 (D2)
- Italy: Liga I (P1)
- Portugal: Series A (I1) and Series B (I2)

2. HomeTeam/AwayTeam - 272 unique values

- D1: 24 teams
- D2: 30 teams
- E0: 26 teams
- E1: 39 teams
- E2: 43 teams

- E3: 41 teams
- I1: 29 teams
- I2: 41 teams
- P1: 23 teams
- SP1: 27 teams
- SP2: 42 teams

3.2. Methods:

Logistic Regression Model (Comprehensive): We employed a comprehensive logistic regression model as our primary analytical method. This model was built using the glm function in R, specifying a binomial distribution with a logit link function. It included a broad set of predictors such as average odds for home and away wins (AvgH, AvgA), goal differences (HGD, AGD), and other match statistics to capture a wide range of factors potentially influencing match outcomes. This approach allowed us to evaluate the combined effect of multiple predictors on the likelihood of a home team win.

Random Forest Model: A Random Forest model was also applied to handle complex interactions and non-linear relationships between predictors more effectively. This method, implemented via the randomForest package in R, provides a robust alternative to logistic regression by aggregating decisions from multiple decision trees built on subsets of the data and features. It helps improve prediction accuracy and robustness against overfitting.

Simplified Logistic Regression Model: In addition to the comprehensive models, we utilized a simplified logistic regression model focusing solely on the most direct predictors of match outcomes—AvgH and AvgA. This model aimed to assess the straightforward impact of betting odds on match results, offering a clear perspective on the predictive power of these key variables in isolation.

3.3. Evaluation:

The dataset was divided into training and testing sets based on match dates rather than a fixed ratio. The training set includes matches before July 1, 2023, totaling 17,331 games. The testing set comprises matches from July 1, 2023, onwards, with 3,570 games. This split ensures that the model is trained on historical data and tested on future, unseen data, mimicking real-world prediction scenarios. This approach avoids random sampling to maintain the chronological integrity crucial for time-series analysis in sports.

Accuracy: Measures the proportion of total correct predictions (both wins and losses) relative to all predictions made. It provides a quick snapshot of model efficacy but does not account for class imbalances or the costs of different types of errors.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): This metric evaluates the model's ability to discriminate between the classes at various threshold settings. The ROC curve plots the true positive rate against the false positive rate, and the AUC provides a single value summarizing overall performance across all thresholds.

Precision and Recall: Precision measures the accuracy of positive predictions (proportion of true positives among all positive predictions), and recall (sensitivity) measures the ability to find all positive instances (proportion of true positives among actual positives). These metrics are particularly useful in scenarios where the costs of false positives and false negatives vary.

F1 Score: The harmonic mean of precision and recall, providing a balance between the two. It is especially useful when classes are imbalanced, as it maintains a balance between the precision and the recall.

Confusion Matrix: A table used to describe the performance of a classification model on a set of test data for which the true values are known. It breaks down the predictions into four outcomes—true positives, false positives, true negatives, and false negatives—providing detailed insight into the types of errors made by the model.

Cross-Validation (CV) Scores: In addition to single-measure performance metrics, cross-validation scores are calculated by training and validating the model on different data set partitions. This method helps ensure the model's performance is stable and reliable across different data subsets, reducing the likelihood of model overfitting and providing a robust measure of its predictive power. Typically, metrics like accuracy, precision, recall, and F1 scores are reported for each fold, and their average provides a comprehensive view of the model's overall effectiveness.

4. **Results:**

As part of our exploratory data analysis, we analyzed the min, first quarter, median, mean, third quarter, and max of our numerical variables to get a better understanding of the dataset. Notably, some variables were calculated before the match based on existing variables obtained from the original dataset for more historical information.

Table 2: Descriptive Statistics of Numerical Variables

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Notes
AvgH	1.06	1.85	2.27	2.59	2.85	34.14	
AvgD	1.88	3.23	3.43	3.67	3.76	15.98	
AvgA	1.08	2.51	3.2	3.84	4.28	37.57	

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Notes
Avg.2.5	1.12	1.76	1.98	1.99	2.17	3.95	
Avg.2.5.1	1.23	1.67	1.82	1.91	2.05	6.01	
AHh	-3.25	-0.5	-0.25	-0.25	0	2.75	
HGD	-61	-7	0	-0.1	6	72	Calculated pre-match
HPts	0	11	23	25.71	38	98	Calculated pre-match
AGD	-60	-6	0	0.12	6	72	Calculated pre-match
APts	0	11	23	25.87	38	98	Calculated pre-match
HLP	1	6	12	11.63	17	24	Calculated pre-match
ALP	1	6	12	11.61	17	24	Calculated pre-match
avgHG	0	1	1.23	1.27	1.54	6	Calculated pre-match
avgAG	0	1	1.24	1.28	1.55	8	Calculated pre-match
avgHST	0	3.39	4	4.04	4.69	14	Calculated pre-match

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Notes
avgAST	0	3.42	4	4.08	4.73	15	Calculated pre-match

4.1. Logistic Regression Model (Comprehensive):

Training Dataset Performance: The comprehensive logistic regression model was trained using a broad array of predictors and showed a stable convergence in terms of coefficient significance and overall fit. During the training phase, the model demonstrated a good ability to capture the underlying patterns within the data, with key predictors like AvgH, AvgA, and league positions showing a significant impact on the match outcomes.

Testing Set Outcomes: On the testing set, the logistic regression model achieved an accuracy of 0.63, an AUC-ROC of 0.69, a precision of 0.63, a recall of 0.82, and an F1 score of 0.71. These results highlight the model's decent discriminatory ability and balanced performance between precision and recall, making it a reliable tool for predicting match results in unseen data.

4.2. Random Forest Model:

Training Dataset Performance: With its ensemble approach, the Random Forest model was particularly effective in managing the complex interactions and nonlinear relationships among the predictors. The out-of-bag (OOB) error estimate provided during training offered an early indication of how well the model might perform on unseen data, which was promisingly low.

Testing Set Outcomes: This model slightly outperformed the logistic regression in accuracy with a score of 0.64 and maintained competitive metrics across the board: AUC-ROC was slightly lower at 0.68, precision was highest at 0.64, while recall was slightly under at 0.81, and the F1 score matched the logistic regression at 0.72. These metrics underscore the Random Forest's strength in handling varied data structures and providing robust predictions.

4.3. Simplified Logistic Regression Model:

Training Dataset Performance: The simplified logistic regression model, using only AvgH and AvgA as predictors, was straightforward but effective in capturing the influence of betting odds on match results. The model’s performance on the training dataset was adequate, with clear indications that these predictors are significant, although the overall fit was less robust compared to the comprehensive model.

Testing Set Outcomes: On the testing dataset, the simplified model's accuracy was slightly lower at 0.62, with an AUC-ROC of 0.65, precision of 0.61, recall the highest at 0.84, and an F1 score of 0.72. The high recall indicates that the model is particularly good at identifying positive (win) instances but at the cost of catching more false positives, as reflected in its lower precision.

4.4. Key Findings from the Validation Process:

Each model provided valuable insights into different aspects of the predictive task. The comprehensive logistic regression offered a balanced approach with good overall metrics. In contrast, the Random Forest excelled in accuracy and precision, suitable for applications where false positives are more costly. The simplified logistic regression highlighted the direct impact of key betting odds, beneficial for quick assessments with limited inputs.

Cross-validation results corroborated these findings, showing that the Random Forest and comprehensive logistic regression models consistently performed well across different subsets of the data, demonstrating their robustness and reliability for broader applications in sports analytics.

Table 3: Performance Comparison

Metric	Logistic Regression	Random Forest	Simplified Logistic Regression
Accuracy	0.63	**0.64**	0.62
AUC-ROC	**0.69**	0.68	0.65
Precision	0.63	**0.64**	0.61
Recall	0.82	0.81	**0.84**
F1 Score	0.71	0.72	**0.72**
CV Accuracy	0.62	**0.63**	0.61

5. **Conclusions and future work**:

This study utilized several statistical and machine learning models to predict the outcomes of football matches based on historical data. The comprehensive logistic regression model provided a balanced approach with an accuracy of 0.63 and an AUC of 0.69, demonstrating its ability to effectively predict match results. The Random Forest model showed a slightly better performance in terms of accuracy (0.64) and precision (0.64), indicating its robustness in handling complex datasets with intricate relationships among variables. The

simplified logistic regression model, while not as robust, highlighted the significant predictive power of betting odds alone, achieving the highest recall of 0.84. The comparative analysis revealed that while the Random Forest model performed slightly better overall, each model has its strengths and could be selected based on the specific needs of the application, such as the trade-off between computational cost and predictive accuracy.

To build on the current findings and enhance the predictive capabilities of our models, several lines of research and development are recommended:

- Incorporation of Additional Data: Including more dynamic variables such as player fitness, recent team performance, and in-season transactions could provide more insights into the likely outcomes of matches. External Factors like weather conditions, player suspensions, and fan support could also be explored for their impact on match results are also helpful.
- Model Enhancement: Exploring more sophisticated ensemble methods like Gradient Boosting or Stacked Models could increase the predictions' accuracy and robustness. Implementing neural networks, especially recurrent neural networks (RNNs) that can process time-series data, might uncover patterns not readily apparent to traditional models.
- Model Deployment and Real-Time Prediction: Developing a system for real-time prediction and updating models with live data during matches could be a significant next step, providing actionable insights instantaneously as match conditions change.
- Cross-League Analysis: Extending the analysis to include multiple leagues from different regions may help generalize the predictive models' effectiveness across various styles of play and competitive levels.
- Methodological Improvements: Applying more rigorous variable selection and model tuning methods, including automated feature selection techniques or hyperparameter optimization algorithms like grid search or Bayesian optimization.

- Impact Studies: Conducting studies on the economic impact of betting predictions on sports betting markets or the potential use of predictive analytics for team management and tactical decisions would provide valuable insights into the practical applications of this research.

Reference:

"2023 Commercial Gaming Revenue Reaches $66.5B, Marking Third-Straight Year of Record Revenue." *American Gaming Association*, 20 Feb. 2024, www.americangaming.org/new/2023-commercial-gaming-revenue-reaches-66-5b-marking-third-straight-year-of-record-revenue/. Accessed 10 Apr. 2024.

Angelini, Giovanni, and Luca De Angelis. "Efficiency of Online Football Betting Markets." *SSRN Electronic Journal*, vol. 35, 2017, pp. 712–721. *2*, doi:10.2139/ssrn.3070329.

"Football Results, Statistics & Soccer Betting Odds Data." *Football Betting - Football Results - Free Bets*, www.football-data.co.uk/data.php. Accessed 20 Mar. 2024.

Gramlich, John. "As More States Legalize the Practice, 19% of U.S. Adults Say They Have Bet Money on Sports in the Past Year." *Pew Research Center*, 14 Sept. 2022, www.pewresearch.org/short-reads/2022/09/14/as-more-states-legalize-the-practice-19-of-u-s-adults-say-they-have-bet-money-on-sports-in-the-past-year/. Accessed 10 Apr. 2024.

Ivankovic, Z., et al. "Analysis of Basketball Games Using Neural Networks." *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*, Nov. 2010, pp. 251–256, doi:10.1109/cinti.2010.5672237.

Lindner, Jannik. "Sports Betting Industry Statistics [Fresh Research] • Gitnux." *GITNUX*, 20 Dec. 2023, gitnux.org/sports-betting-industry-statistics/. Accessed 10 Apr. 2024.

Loeffelholz, Bernard, et al. "Predicting NBA Games Using Neural Networks." *Journal of Quantitative Analysis in Sports*, vol. 5, no. 1, 15 Jan. 2009, doi:10.2202/1559-0410.1156.

Puranmalka, Keshav. *Modeling the NBA to Make Better Predictions*, Massachusetts Institute of Technology, 1 Jan. 1970, dspace.mit.edu/handle/1721.1/85464. Accessed 14 Apr. 2024.

Sibony, John, et al. "Neural Networks Predictive Modeling for Football Betting." *SSRN*, 2 Sept. 2020, papers.ssrn.com/sol3/papers.cfm?abstract_id=3655700. Accessed 12 Apr. 2024.

Song, ChiUng, et al. "The Comparative Accuracy of Judgmental and Model Forecasts of

American Football Games." *International Journal of Forecasting*, vol. 23, no. 3, July 2007, pp. 405–413, doi:10.1016/j.ijforecast.2007.05.003.

"Value of Global Bets on European Soccer 2021." *Statista*, 7 Sept. 2023, www.statista.com/statistics/1263462/value-betting-on-european-soccer/. Accessed 11 Apr. 2024.

Wilkens, Sascha. "Sports Prediction and Betting Models in the Machine Learning Age: The Case of Tennis." *SSRN*, 9 Jan. 2020, papers.ssrn.com/sol3/papers.cfm?abstract_id=3506302. Accessed 13 Apr. 2024.