PREDICTIVE MODELING FOR SPORTS BETTING

THE CASE OF SOCCER

Diep Vu
Yousef Lari

TABLE OF CONTENTS

01 INTRODUCTION

02 RELATED WORKS

03 MATERIALS

04 METHODS

05 RESULTS

06 FUTURE WORK


01
INTRODUCTION

HISTORY OF SPORTS BETTING IN THE U.S.

- Illegal under the Professional and Amateur Sports Protection Act of 1992 (PASPA)
- Later struck down by the Supreme Court later in 2018
- Sportsbooks are legal in 37 states and the District of Columbia
- Online sports betting is also legal in 30 states and Washington D.C

ANNUAL REVENUE

2021

$57.22 billion handle
$4.29 billion revenue

2022

$93.2 billion handle
$7.56 billion revenue

2023

$119.8 billion handle
$10.9 billion revenue

OBJECTIVES AND IMPORTANCE

- Objectives include developing predictive models for soccer match outcomes, identifying significant factors, and comparing statistical methods
- Utilizes comprehensive datasets with pre-match betting odds and aim to revolutionize betting strategies through data-driven insights
- Predictive models in sports betting have significant potential to elevate the industry to new levels of sophistication and success.

02

RELATED WORKS

OVERVIEW OF RELATED WORKS

- Various studies about predictive modeling in sports:
  - American football
  - Basketball
  - European football
  - Horse Racing
  - Tennis
- Various competitions including
  - National leagues
  - International tournaments
  - The Olympics.
- Advancements in neural networks and the integration of external data sources to enhance accuracy

EXAMPLES OF PREDICTIVE MODELS

- Studies have utilized neural network models such as feed-forward, radial basis, probabilistic, and generalized regression neural networks to predict sports outcomes.
- Fusion techniques, including Bayes belief networks and probabilistic neural network fusion, have been explored to improve model performance.
- Artificial Neural Networks (ANNs) to predict outcomes of basketball games by formalizing effects of shots from different court areas as features

INTEGRATION OF EXTERNAL DATA SOURCES

- Incorporated external data sources like player transfer news, social media sentiment analysis, and betting market data to enhance predictive models.
- Examples include using Twitter posts to predict NFL game outcomes and leveraging play-by-play data to develop new features, outperforming models based solely on traditional statistics.

03

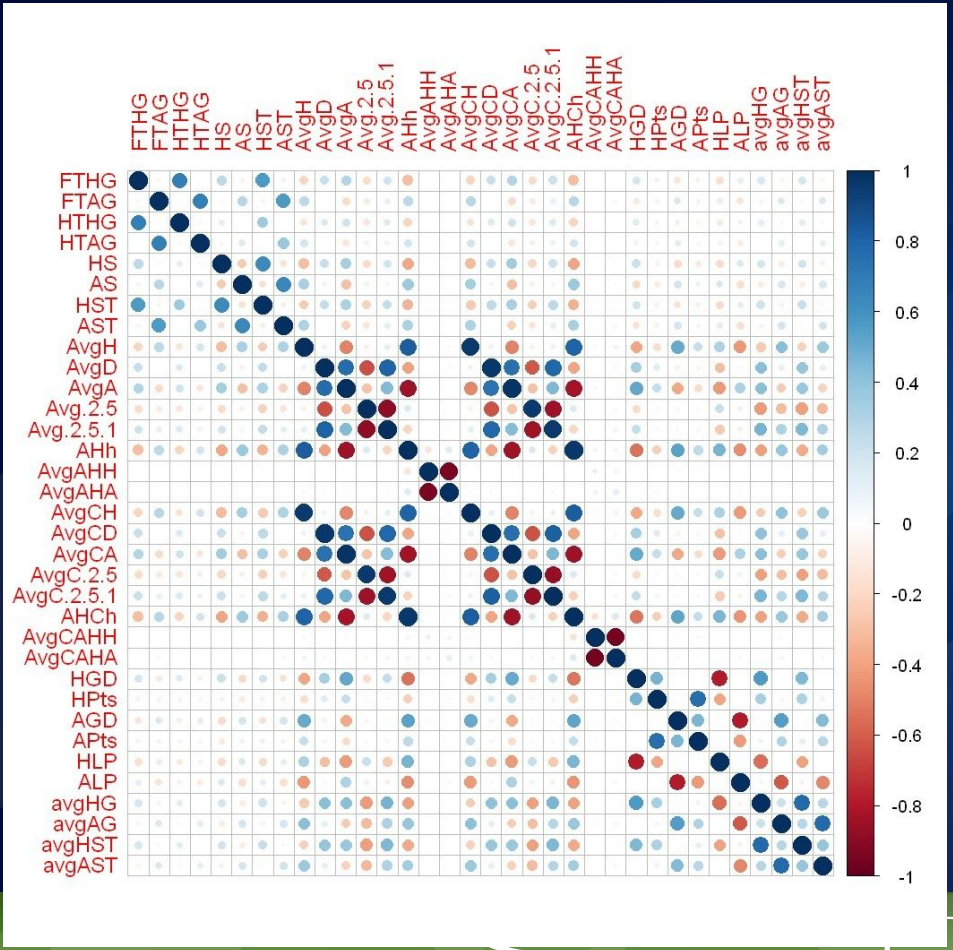
MATERIALS

DATASET

- The data was collected from Football-Data.co.uk and includes 43 files from different leagues in 5 countries (England, Germany, Italy, Spain, and Portugal) spanning 4 seasons from 2020 to 2024.
- Each league season was collected separately and loaded into a dataframe using Pandas in Python.
- The dataset initially contained 20921 matches with 111 variables. However, after preprocessing, only 38 columns with necessary information were retained.

PREPROCESSING

- Preprocessing steps included generating average goals, average shots on target as well as points and ranks for each team before each match based on the number of wins, draws, losses, and goal difference
- Only average betting odds were retained, and null values in the dataset were dropped, resulting in a dataset of 20,901 matches with 40 variables.
- It's important to note that the odds are in decimal format, representing the amount a winning bet would collect on a $1 bet.



- Keep only pre-match variables
- Average odds and closing odds are highly correlated, so closing odds were removed
- Final dataset of **20,901 matches** with **21 variables**

TABLE I. VARIABLE DESCRIPTION

Variable	Description
Div	Categorical, league division
HomeTeam	Categorical, home team
AwayTeam	Categorical, away team
AvgHG	Numerical, average home team goals
AvgAG	Numerical, average away team goals
AvgHST	Numerical, average home team shots on target
AvgAST	Numerical, average 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)
HGD	Numerical, home team goal difference
AGD	Numerical, away team goal difference
HPts	Numerical, home team points
APts	Numerical, away team points
HLP	Numerical, home team league place
ALP	Numerical, away team league place

TABLE 2. NUMERICAL VARIABLES STATISTICS

Variable	Min	Mean	Max
AvgHG	0	1.27	6
AvgAG	0	1.28	8
AvgHST	0	4.04	14
AvgAST	0	4.08	15
AvgH	1.06	2.59	34.14
AvgD	1.88	3.67	15.98
AvgA	1.08	3.84	37.57
Avg>2.5	1.12	1.99	3.95
Avg<2.5	1.23	1.91	6.01
Ahh	-3.25	-0.25	2.75
HGD	-61	-0.10	72
AGD	-60	0.12	72
HPts	0	25.71	98
APts	0	25.87	98
HLP	1	11.63	24
ALP	1	11.61	24

CATEGORICAL VARIABLES

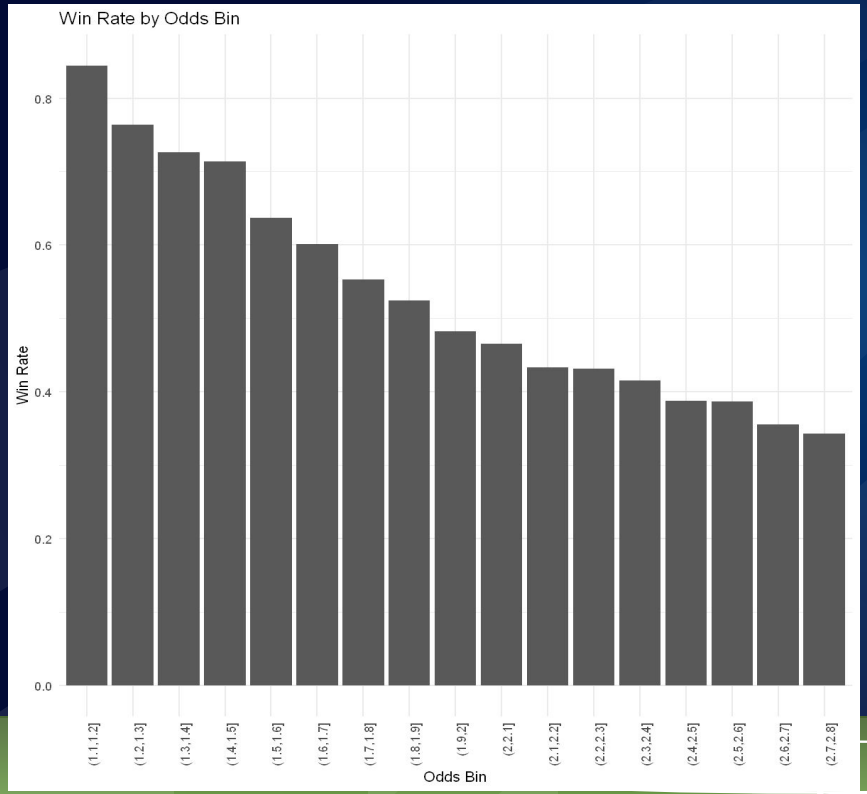
11 LEAGUES

- 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)

272 TEAMS

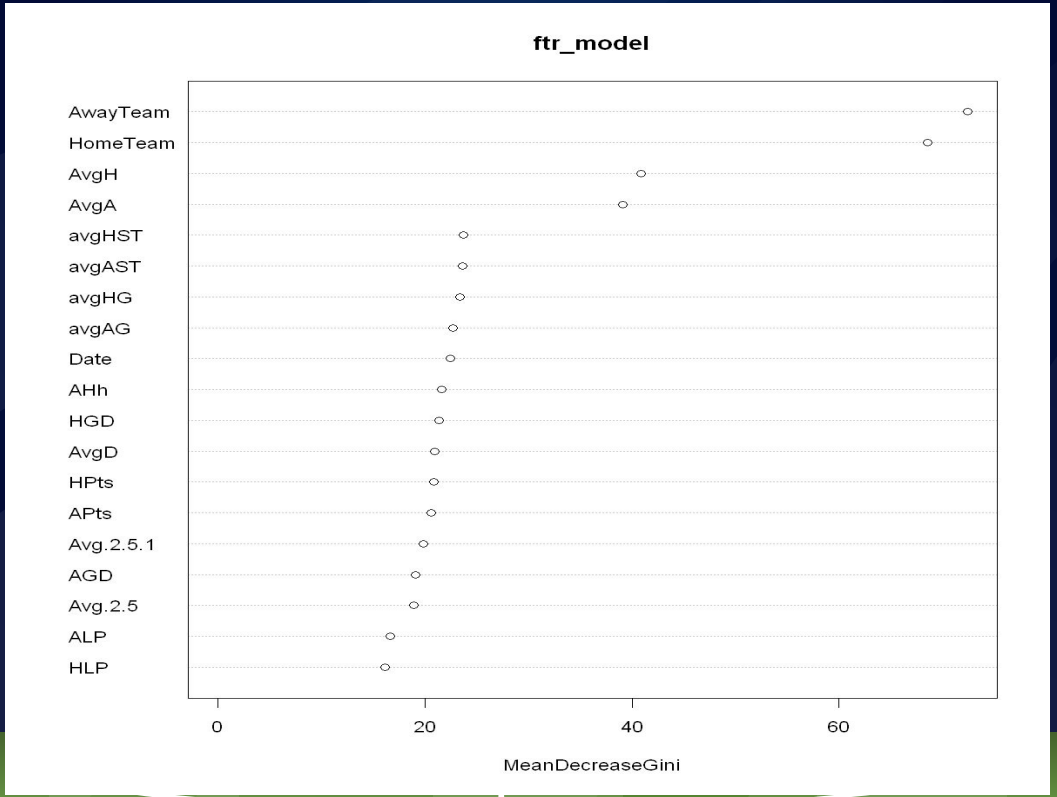
- D1 : 24
- D2 : 30
- E0 : 26
- E1 : 39
- E2 : 43
- E3 : 41
- I1 : 29
- I2 : 41
- P1 : 23
- SP1 : 27
- SP2 : 42

WIN RATE BY ODDS



	Odds_Bin	Win_Rate	Count
1	(1.1,1.2]	0.845	283
2	(1.2,1.3]	0.764	444
3	(1.3,1.4]	0.726	665
4	(1.4,1.5]	0.714	836
5	(1.5,1.6]	0.637	1038
6	(1.6,1.7]	0.602	1315
7	(1.7,1.8]	0.553	1448
8	(1.8,1.9]	0.525	1450
9	(1.9,2]	0.483	1647
10	(2,2.1]	0.466	1808
11	(2.1,2.2]	0.433	1846
12	(2.2,2.3]	0.432	1807
13	(2.3,2.4]	0.415	1803
14	(2.4,2.5]	0.388	1804
15	(2.5,2.6]	0.387	1554
16	(2.6,2.7]	0.356	891
17	(2.7,2.8]	0.343	213

MEAN DECREASE GINI



04

METHODS

ANALYTICAL METHODS

1. RANDOM FOREST

- **Purpose:** Classify match outcomes as 'Won' or 'Lost'.
- **Method:** Employ the `randomForest` package to train a model excluding the `FTR` variable to avoid data leakage.
- **Evaluation:** Assess model performance using accuracy, F1 score, and a confusion matrix derived from test data.

2. GENERALISED LINEAR (GLM)

- **Purpose:** Compare predictive performance with a different statistical approach.
- **Method:** Fit a logistic regression model using the glm function with predictors like HomeTeam, AwayTeam, AvgH, and AvgA.
- **Evaluation:** Utilize ROC curves, AUC scores, and cross-validation to validate model accuracy and robustness.

3. GLM WITH INTERACTION TERMS

- **Purpose:** Explore the effect of interactions between teams on match outcomes.
- **Method:** Extend the GLM to include interaction terms between `HomeTeam` and `AwayTeam`, alongside other predictors.
- **Evaluation:** Review model effectiveness through modified accuracy, AUC, and the clarity of model predictions compared to previous models.

05

RESULTS

METHOD EVALUATION

Model Type	Accuracy	AUC	Precision	Recall	F1 Score
Random Forest	64.27%	0.6377	68.41%	67.73%	63.75%
GLM (without interactions)	65.79%	0.7193	66.52%	76.07%	70.98%
GLM (with interactions)	60.53%	0.6042	64.36%	61.58%	62.94%

EXPLANATION OF METRICS

- **Accuracy:** Percentage of total correct predictions.
- **AUC:** Area Under the ROC Curve, a measure of the ability of the model to avoid false classifications.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall:** The ratio of correctly predicted positive observations to all observations in actual class.
- **F1 Score:** The weighted average of Precision and Recall.

06

CONCLUSION AND FUTURE WORK

FUTURE MODEL IMPROVEMENTS

- **Model Ensemble Techniques:** Combine predictions from multiple models through techniques like stacking or blending to potentially improve prediction accuracy and reduce variance.
- **Advanced Machine Learning Models:** Explore more sophisticated machine learning techniques, such as Gradient Boosting Machines (GBMs) or Neural Networks, which might capture nonlinear relationships and interactions more effectively.
- **Temporal Analysis:** Incorporate time-series analysis to account for changes in team performance over seasons or to understand trends and cycles in the data.
- **Hyperparameter Optimization:** Utilize methods like grid search or random search to fine-tune model parameters, potentially improving model outcomes significantly.
- **Cross-Validation Schemes:** Experiment with different cross-validation strategies, such as stratified or time-series splits, to ensure the model is robust and generalizes well to unseen data.

FUTURE WORK

- Refining Feature Engineering:

Incorporate additional contextual factors like division and statistics of the last 5 games.

- Exploring Alternative Machine Learning Algorithms:

Investigate the use of neural networks and other algorithms.

- Addressing Class Imbalance:

Implement techniques like oversampling or synthetic data generation to improve predictive performance and robustness in forecasting football match outcomes.

LIST OF REFERENCES

- American Gaming Association. “2023 Commercial Gaming Revenue Reaches $66.5B, Marking Third-Straight Year of Record Revenue.” American Gaming Association, www.americangaming.org/new/2023-commercial-gaming-revenue-reaches-66-5b-marking-third-straight-year-of-record-revenue/.
- 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/.
- Gitnux. “Sports Betting Industry Statistics.” Gitnux, www.gitnux.org/sports-betting-industry-statistics/.
- Statista. “Value of Betting on European Soccer Worldwide in the 2020/2021 Season, by League.” Statista, www.statista.com/statistics/1263462/value-betting-on-european-soccer/.
- Loeffelholz, Bernard, et al. "Predicting NBA Games Using Neural Networks." Journal of Quantitative Analysis in Sports, vol. 5, no. 1, 2009, https://doi.org/10.2202/1559-0410.1156.
- Song, ChiUng, et al. "The Comparative Accuracy of Judgmental and Model Forecasts of American Football Games." International Journal of Forecasting, vol. 23, no. 3, 2007, pp. 405-413, https://doi.org/10.1016/j.ijforecast.2007.05.003.
- Sibony, John, et al. "Neural Networks Predictive Modeling for Football Betting." SSRN, 19 July 2020, https://ssrn.com/abstract=3655700 or http://dx.doi.org/10.2139/ssrn.3655700.

QUESTIONS?