Soccer Match Prediction

Christopher Semaan &
Bardia Parmoun

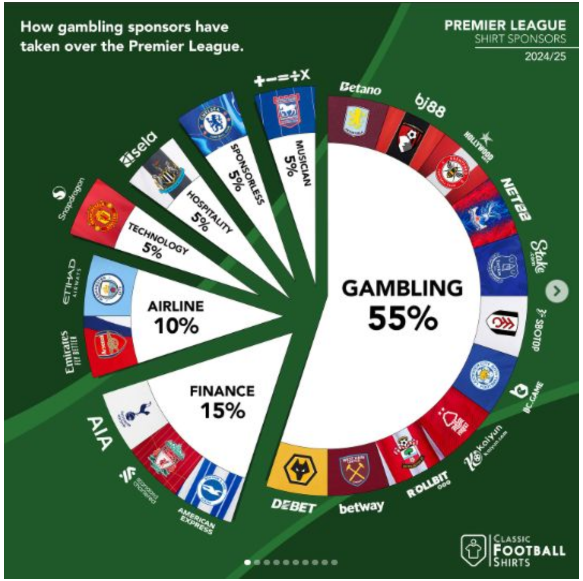
Introduction

Motivation

- Premier League: **900 million** homes worldwide [1]
- **4% YOY** growth in new markets [1]
- Record breaking viewership in 2024 [1]

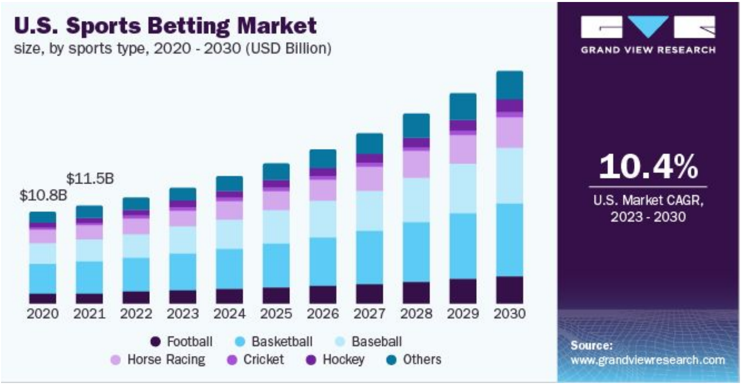
Growth of Sports Gambling

- **11 Clubs** sponsored by gambling [2]



- **10.4%** Annual Growth Rate
- **83.65 billion USD** worldwide

[3]



Objective

- Predict Match Results **(Home Win, Draw, Away Win)**

- Maximize **profit** based on each match’s given odds

Dataset Search Requirements

- **Real** match results
- **Odds** for HomeWin, AwayWin, and Draw
- **Pre-match** data

Background

What are odds?

- A payout ratio [4] -> **i.e** 5:1 means earn 5 times the bet amount
  - Often made by professional odds makers
  - **Inversely** correlated with likelihoods

- **Lowest odd -> most likely outcome**

- **Profit calculation using odds:**

$$Profit = \sum [outcome\_odds \cdot (predicted\_outcome == actual\_outcome) - 1]$$

Possible approaches

- **Traditional odds making**
  - Using insider information
  - Historical analysis

- **Machine Learning**
  - **Simple models:** Support Vector Machines, random forest
  - **Deep learning:** FNN, RNN, LSTM

Survey of existing solutions

- **Odds makers:** traditional odds makers (i.e **Bet365**) [5]
  - **Performance:** usually has an accuracy of **~55%** on average
- **Voting model:** FNN and Random Forest [6]
  - **Features:** goals, fouls, cards, penalty kicks, own goals, free kicks, own goals …
  - **Performance:** only achieved an accuracy of **46.6%**
- **Sequential model:** RNN and LSTM [7]
  - **Features:** win/loss streaks (3 and 5 matches), past 4 results, goals scored, …
  - **Performance:** the model achieved an accuracy of **81.75%**
  - Does not predict draws and outperforms other models by **~20% -> maybe not reliable**

Survey of existing solutions

- **Decision Trees:** Random Forest and Gradient Boosting [8]
  - **Features:** rank difference, goal difference, goal per rank differences, …
  - **Performance:** achieved an accuracy of **71.72%**
- **Deep Multi-layer Learning:** 5 layer FNN [9]
  - **Features:** shots, shots on target, corners, fouls, first-half and total goals, …
  - **Performance:** only achieved an accuracy of **61.14%**
- **Stacking Model:** CNN, SVM, Logistic Regression, and Random Forest [10]
  - **Features:** total time, cards, fouls, corners, shots on target, humidity, temperature, wind speed, weather conditions, …
  - **Performance:** the model achieved an accuracy of **62.6%** and a F1 score of **59.2**

Selected Approach

- **Multi-layer Feedforward Neural Network (FNN) - 5 layers**
  - FNN was used in the **majority** of the existing approaches
  - Deep learning is preferred due to **complexity** -> no random forest
  - Historical data is already **aggregated** -> RNN and LSTM not needed
  - All the fields are simple **numerical values** -> CNN not needed
  - Lots of features -> need to optimize **training time** so no transformers
- **Expectations:**
  - Accuracy ranges between low **40-60%** -> we aim to get **~50%**

System Description

Dataset Detailed

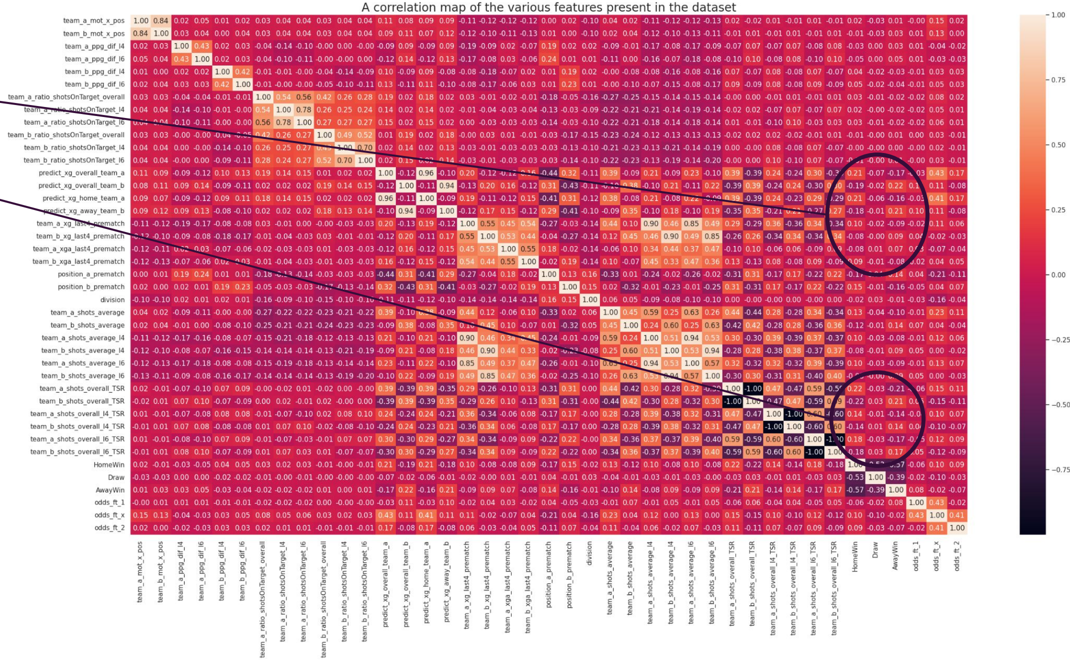
Columns	Type
team_a_shots_average **team_a_shots_overall_TSR** **team_a_ratio_shotsOnTarget_overall** **team_a_ppg_dif_l6** **position_a_prematch** **...**	Prematch (Historical)
predict_xg_overall_team_a **predict_xg_home_team_a** **...**	Predictions
odds_ft_1 **profit_1** **...**	Betting information

45 Columns

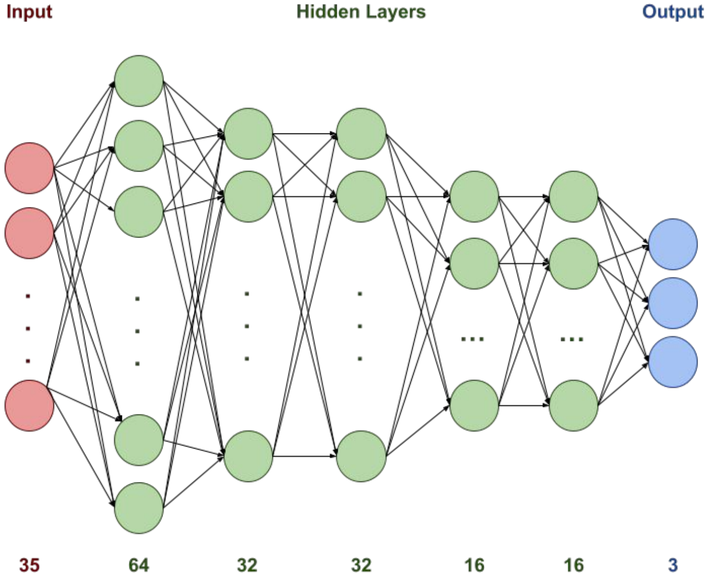
57819 Matches

Analyzing the columns

- **Important columns:**
  - xG vs Results
  - Total Shot Ratio vs Results
- **Ignored Columns:**
  - **Extra Information:**
    - 'Profit_1'
    - 'Profit_X'
    - 'Profit_2'
    - Profit metric was used instead
  - **Redundant Features:**
    - 'team_a_ratio_shotsOnTarget_l6'
    - 'team_a_ratio_shotsOnTarget_l4'
    - 'team_b_ratio_shotsOnTarget_l6'
    - 'team_b_ratio_shotsOnTarget_l4'
    - **Already correlated with TSR!**



Model Description



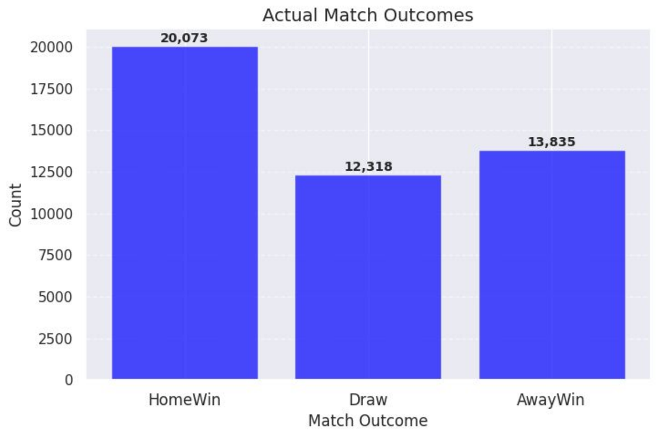
Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	2,048
batch_normalization (BatchNormalization)	(None, 64)	256
leaky_re_lu (LeakyReLU)	(None, 64)	0
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2,080
batch_normalization_1 (BatchNormalization)	(None, 32)	128
leaky_re_lu_1 (LeakyReLU)	(None, 32)	0
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 32)	1,056
batch_normalization_2 (BatchNormalization)	(None, 32)	128
leaky_re_lu_2 (LeakyReLU)	(None, 32)	0
dropout_2 (Dropout)	(None, 32)	0
dense_3 (Dense)	(None, 16)	528
batch_normalization_3 (BatchNormalization)	(None, 16)	64
leaky_re_lu_3 (LeakyReLU)	(None, 16)	0
dropout_3 (Dropout)	(None, 16)	0
dense_4 (Dense)	(None, 16)	272
batch_normalization_4 (BatchNormalization)	(None, 16)	64
leaky_re_lu_4 (LeakyReLU)	(None, 16)	0
dropout_4 (Dropout)	(None, 16)	0
dense_5 (Dense)	(None, 3)	51

Our Algorithm

- **Feedforward Network**
  - Input: 35 features
  - Layer **1: 64** Neurons
  - Layer **2, 3: 32** Neurons
  - Layer **4, 5: 16** Neurons
  - **Output** layer: **3** Neurons

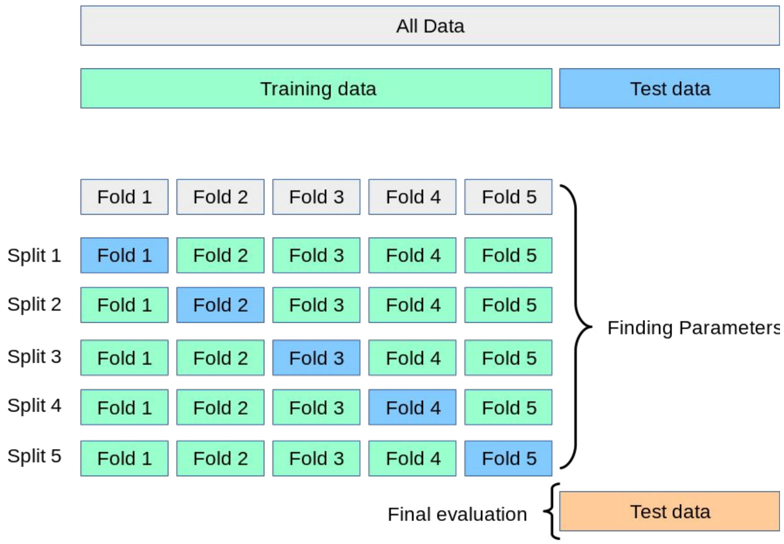
- **All layers**
  - L2 Regularization: **0.2** Factor
  - Batch Normalization
  - Leaky ReLU: **-0.01** slope
  - Dropout: **0.2** rate
  - Balanced class weights for **Draw** & **AwayWin**

- **Adam Optimizer**
  - Adaptive learning rate
    - **0.1** reduction factor
    - **3** epoch patience factor



Testing and Training

- **Training Splits**
  - 5-Fold Cross Validation
    - **70%** Training on Cross Validation
    - **10%** Validation
    - **20%** Testing
  - Final Validation
    - **90%** Training
    - **10%** Validation
- **Training Parameters**
  - **50** Epochs
  - **32** Batch Size



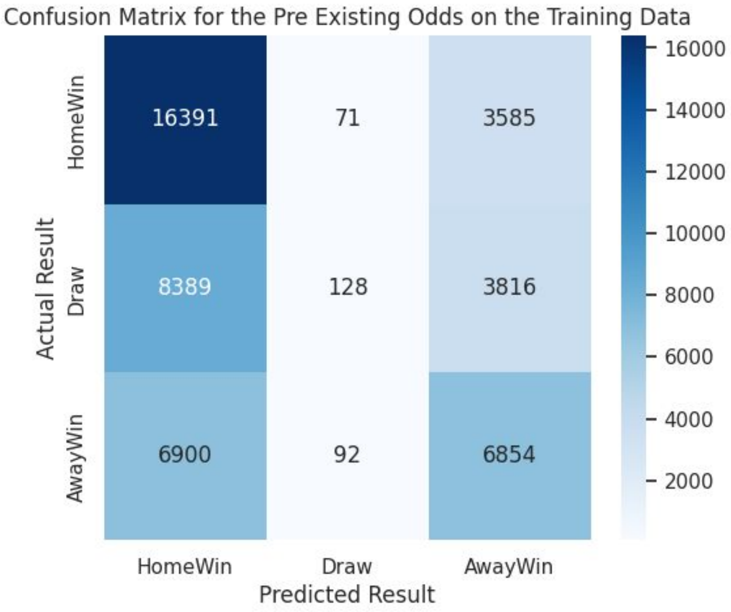
Numerical Results

Existing Odds on the Training Data

- The existing odds have an accuracy of **50.56%** on **all** the training data

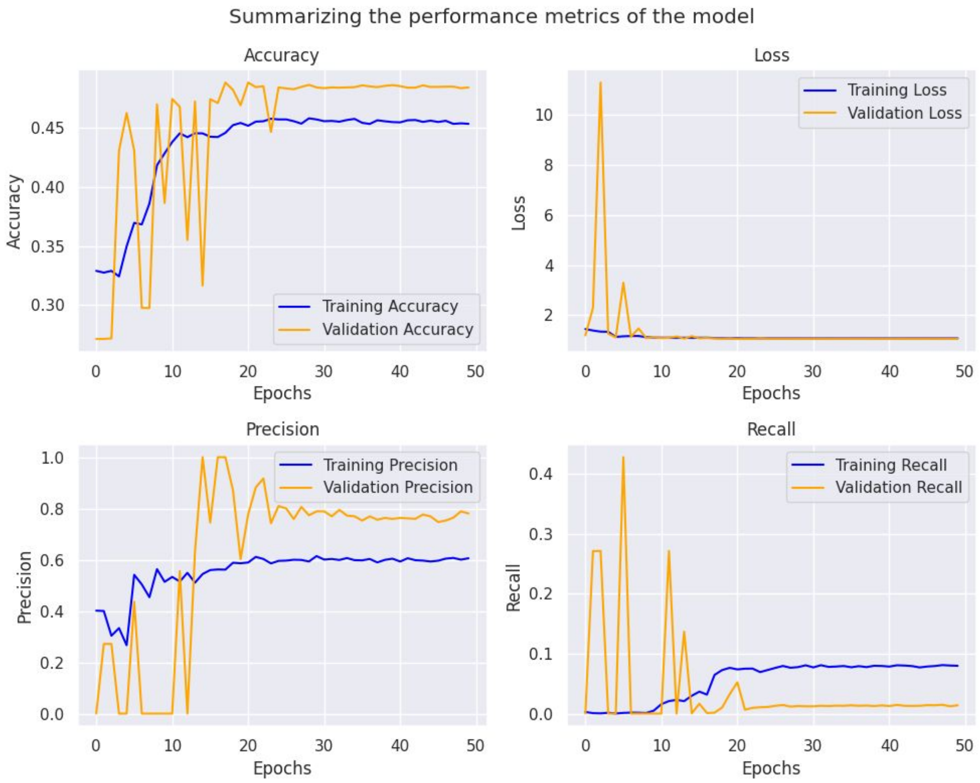
- **Profit Metric:**
  The odds results in a net profit of **$-2401.60**

- **Heavily biased against draws!**



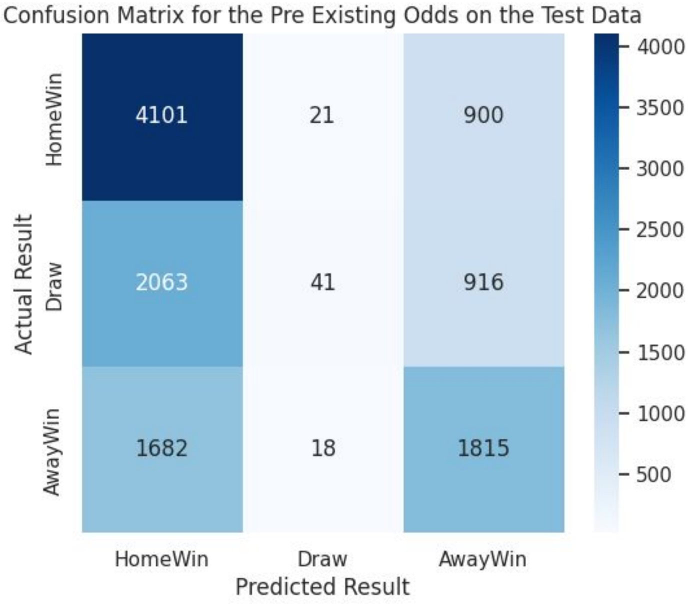
Model Performance during Cross Validation

- **F1 score:** 46.28 ± 1.58
- **Model’s net profit:** $-441.47 ± $122.94
- **Odd’s net profit:** $-505.19 ± $40.79
- **Maximum accuracy:** 48.43%



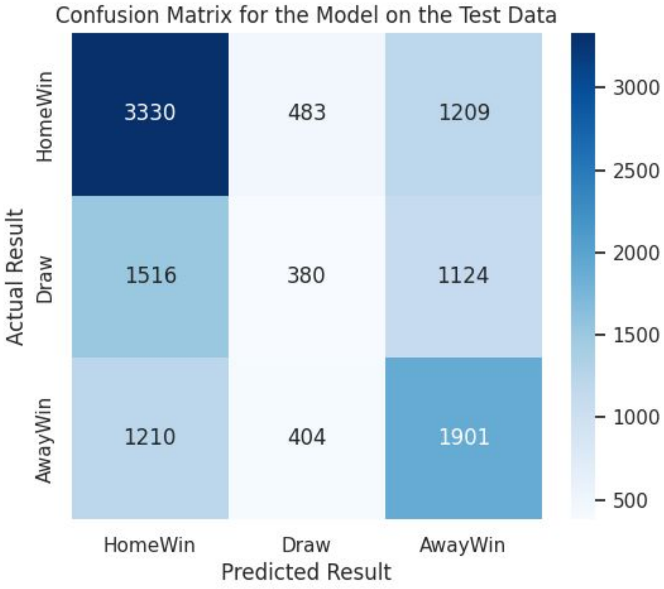
Performance on the Blind Test Data

Existing Odds



Accuracy: 51.54%
Net Profit: $-330.72

The Model



Accuracy: 48.55%
Net Profit: $-372.68

Findings

- **Performance:**
  - Performs **closely** to existing odds!
  - Outperforms random (33%) by **15%**!
  - Better than some of the existing models!

- **less bias** towards **draws** -> compared to the odds
- Recall is not good → the model is too **strict**!
- Accuracy does not cross 50% -> need more **features**

Conclusion

Conclusions

- **Summary:**
  - Our best statistical run reached **48.43%** accuracy
  - Our results were more balanced towards Draws and Away Wins than the Odds makers
- **Suggestions:**
  - More samples would make the model more consistent and accurate
  - Include player specific features or more diverse team factors

Final Thoughts

- Soccer is **NOT** science!
- Lots of hidden factors!

- **Example: Manchester City vs Cardiff City [5]**
  - Bet365 odds for Cardiff win were **9:1**
  - Manchester City had **74%** possession
  - Manchester City had **10** shots on target as opposed to Cardiff's **4**
  - Yet, Cardiff **won** 2-0!

References

[1] The numbers that show this has been a season like no other. Available: https://www.premierleague.com/news/4027356. Accessed: 2025-03-26.

[2] @classicfootballshirts. Gambling sponsors in the premier league 2024/25. Available: https://www.instagram.com/classicfootballshirts/p/C-slVewK-_a/?img_index=7. Accessed: 2025-03-26.

[3] Grandview Research. Sports betting market size, share & trends analysis report by platform, by betting type (fixed odds wagering, exchange betting, live/in-play betting, esports betting), by sports type, by region, and segment forecasts, 2023 - 2030. Available: https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report. Accessed: 2025-03-26.

[4] Shehryar Sohail. Sports betting odds: How they work and how to read them. Available: https://www.investopedia.com/articles/investing/042115/betting-basics-fractional-decimal-american-moneyline-odds.asp. Accessed: 2025-03-28.

[5] socceranalytics. Predictions in soccer: Getting things right more often than wrong. Available: https://socceranalytics.org.uk/predictions-in-soccer/. Accessed: 2025-03-28.

References

[6] Sherif Saad Haytham Elmiligi. Predicting the outcome of soccer matches using machine learning and statistical analysis. Available: https://ieeexplore.ieee.org/document/9720896. Accessed: 2025-03-27.

[7] Sarika Jain. Soccer result prediction using deep learning and neural networks. Available:https://www.researchgate.net/publication/349272309_Soccer_Result_Prediction_Using_Deep_Learning_and_Neural_Networks. Accessed: 2025-03-28.

[8] Xiangkun Meng. Soccer match outcome prediction with random forest and gradient boosting models. Available: https://www.researchgate.net/publication/378355415_Soccer_match_outcome_prediction_with_random_forest_and_gradient_boosting_models. Accessed: 2025-03-27.

[9] Sergei Bezobrazov Sergei Anfilets. Deep multilayer neural network for predicting the winner of football matches. Available: https://www.researchgate.net/publication/342416626_DEEP_MULTILAYER_NEURAL_NETWORK_FOR_PREDICTING_THE_WINNER_OF_FOOTBALL_MATCHES. Accessed: 2025-03-28.

[10] Eugene Li Albert Wong. A predictive analytics framework for forecasting soccer match outcomes using machine learning models. Available: https://www.sciencedirect.com/science/article/pii/S2772662224001413. Accessed: 2025-03-28.

Thank you!