15.095 Machine Learning Under A Modern Optimization Lens

An Optimal Betting Framework for the 2022 FIFA World Cup

Yen Hann Yoo | Shannan Liu

1.0 Introduction and Overview

Football is a sport spectated by billions annually. According to FIFA (2018), the 2018 FIFA World Cup had over 3.5 billion viewers, with cumulative betting valued at $155 billion. However, the optimal betting strategy is often unclear for maximizing returns due to the tradeoff between risk and reward. For example, betting on Argentina to win against Saudi Arabia will result in smaller winnings but a higher probability of winning (i.e., lower risk, lower reward). Conversely, betting on Saudi Arabia to win against Argentina will have higher returns but lower probability (i.e., higher risk, higher reward). In this project, we explore the use of Random Forest, XGBoost, Light Gradient Boosting Machine, and Logistic Regression to prescribe an optimal betting strategy that maximizes expected betting returns on the group stages. The XGBoost model will then be used to predict the knockout stages.

2.0 Data Gathering and Pre-Processing

2.1 International Matches

A dataset containing ~4000 international matches played between 2004 to 2022 forms the primary dataset. Predictors are presented as pairs (home and away team). These include:

- Team continent (North America, South America, Europe, etc.)
- FIFA team rankings and total points
- Tournament type (FIFA World Cup, Qualification, UEFA Euro, etc.)
- City and country where match was played
- Neutral location (home advantage or not)
- Best goalkeeper rating (rating of the highest rated goalkeeper)
- Average defensive, midfield, and offensive rating

2.2 Betting Odds Data for Group Stages

At the time of writing, a betting framework could not be applied to the knockout stages because the World Cup has not concluded. Therefore, the betting framework was only applied to the group stages. The betting odds data for each group stage match were obtained from www.oddsportal.com. The odds on this site were enumerated in an “American” format. American odds are represented in terms of expected wins around a $100 wager. For instance, if the odds are -130, then betting $130 would result in $100 earnings on the desired outcome. If the odds are 150, then betting $100 would result in $150 earnings on the desired outcome.

For the prescription framework, the odds were converted into expected earnings for every dollar bet on an outcome. The conversion was performed as follows

$$if\ (Betting\ Odds)\ <\ 0,\ then\ earnings\ on\ betting\ \$1\ = \frac{-100}{Betting\ Odds)}$$
$$if\ (Betting\ Odds)\ \geq\ 0,\ then\ earnings\ on\ betting\ \$1\ = \frac{Betting\ Odds}{100}$$

This conversion was completed for every outcome of each match so that we could later compute each match's most profitable outcome for a bettor to invest in.

2.3 Weather Data, City Populations

To increase model accuracy, the average temperature, in degrees celsius, of a city on the date a match was played was integrated into the dataset. It was hypothesized that this information would be helpful for making predictions because it would reflect an aspect of a match's environment affecting teams' performances. For instance, if a team from Europe, which is generally colder than South America, played in warmer climates, then their performance may be affected. Under this assumption, weather data was included into the dataset for predictive purposes.

2.4 Data Preprocessing

After collecting the data, we dropped the following variables that would not be useful for general prediction purposes or that could cause leakage in the modeling process: 'date', 'tournament', 'city', 'country', 'neutral_location', 'lat', 'lng', 'home_score', 'away_score'. Subsequently, we converted all categorical variables into dummy variables. Then, we scaled the data via standardization.

This processing procedure was applied to the testing set, which we evaluated our models on, as well as the dataset with group stage games with the additional caveat that we added any missing dummy variable columns to the testing and group stage datasets.

3.0 Modeling, Results, and Discussion

In the modeling process, our target variable was a match's outcome. This could be separated into three categories, "Win", "Loss", "Draw" with respect to the "home team" of an observation. The variables we used to train models were described above, including FIFA team rankings, points, continents where teams are from, and weather data.

Each model was parameterized via cross validation and then evaluated on several metrics with an unseen portion of international match data. As can be seen from Table 1, the cross-validated XGBoost model performed the best. Hence it was used to make group-stage match predictions for the world cup. The out-of-sample results of our models are shown in this table.

Table 1 | Model Performances

	Random Forest	XGBoost	LGBM	Logistic Regression
AUC	**0.8587**	**0.9119**	**0.903**	**0.73**
Accuracy	**0.67**	**0.78**	**0.78**	**0.59**
Precision	**0.76**	**0.77**	**0.77**	**0.55**
Recall	**0.59**	**0.75**	**0.76**	**0.52**
F1-Score	**0.58**	**0.76**	**0.76**	**0.52**

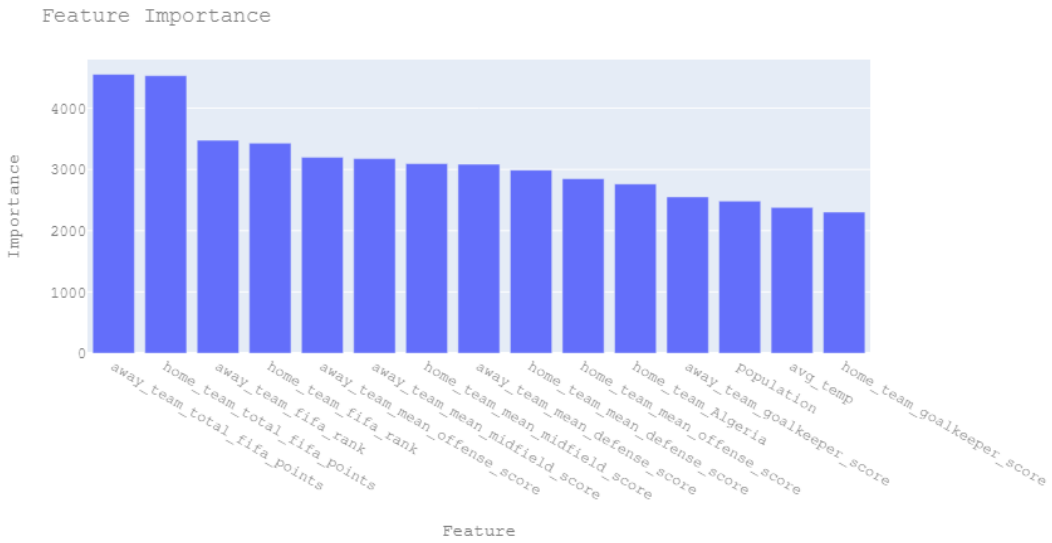

Figure 1 | Feature Importance Plot for XGBoost Model

This feature importance plot shows that the most important features for the model's predictions are the home and away team's total FIFA points. FIFA points are computed by adding the average number of points a team has gained during matches in the last 12 months to the average number of points they gained from games older than 12 months, which is subject to some level of depreciation each year. Hence, a team's FIFA points are an up-to-date measurement of a team's recent performance. As a result, it makes sense that the model considers these variables significant predictors. Because a team's FIFA points and FIFA rank are correlated, it would make sense that a team's FIFA rank is also an important variable to the XGBoost model.

From the feature importance plot, we can also see the significance of a team's offense, midfield, defense, and goalkeeping play in the model's predictions. Notably, the model considers midfield scores the most important, followed by defense, offense, and then goalkeeping. Although the differences in importance rating are small, this suggests that a team's midfield performance is more important to a match's outcome than offense, defense, or goalkeeping performance.

Following the strong results of the model, we proceeded to make predictions on the world cup group-stage matches. Once the predictions were made, we then developed a prescription framework for betting where we bet on the match outcome with the highest expected return. This is detailed in the next section.

5.0 Betting Framework for Group Stages

Figure 2 | 2022 FIFA World Cup Group Stages

5.1 Optimization Formulation

Six matches are played per group, which equates to a total of 48 group stage matches. The goal of the optimization framework was to maximize expected profit from betting, given data on earnings per dollar (R_{ij}) and probabilities (p_{ij}) determined by the XGB model for each outcome *j* across all matches *i*. This was formulated as follows:

Sets

- Matches: i = 1, ..., 48
- Outcomes: j = 1, 2, 3 (1 = Home win, 2 = Home draw, 3 = Home loss)

Data

- Probabilities of each group stage match *i* and outcome *j* from model: p_{ij}
- Earnings per dollar on a group stage match *i* and outcome *j*: R_{ij}

Decision Variables

- x_{ij}: Betting amount on match *i* for outcome *j*
- z_{ij}: 1 if bet on match *i* and outcome *j*, 0 otherwise

Objective Function

Maximize expected profits

$$max \sum_{i=1}^{48} \sum_{j=1}^{3} (p_{ij} R_{ij} x_{ij}) - \sum_{i=1}^{48} \sum_{j=1}^{3} x_{ij}$$

Constraints

1) Can only bet on one outcome *j* at most for each match *i*

$$\sum_{j=1}^{3} z_{ij} \leq 1 \qquad \forall i$$

2) The value of a bet cannot exceed 5% (1/20) of the budget

$$x_{ij} \leq \frac{B}{20} z_{ij} \qquad \forall i, \forall j$$

3) Total value of all bets cannot exceed available budget, B

$$\sum_{i=1}^{48} \sum_{j=1}^{3} x_{ij} \leq B$$

4) Risk spreading over at least *K* matches

$$\sum_{i=1}^{48} \sum_{j=1}^{3} z_{ij} \geq K$$

5) If $z_{ij} = 1$, a betting amount x_{ij} must be made

$$x_{ij} \geq z_{ij} \qquad \forall i, \forall j$$

5.2 Benchmarking: "God's Eye" Betting Strategy

A benchmark target was developed: "God's Eye," where perfect knowledge of each match and outcome is assumed. In the example of *Argentina v. Saudi Arabia*, "God's Eye" would predict Saudi Arabia to win and allocated a betting amount (if any) to a Saudi Arabia win. The full formulation of "God's Eye" can be found in Appendix A. The profits were computed following the conclusion of the group stages as follows:

$$If\ prediction = actual\ outcome\text{: } Profit\ for\ match\ i = x_{ij}R_{ij} - x_{ij}$$
$$If\ prediction \neq actual\ outcome\text{: } Profit\ for\ match\ i = -x_{ij}$$

Table 2 | Comparison of benchmarks against betting framework (total profit)

Total Bet Amount : **$5000**	**Model (XGB)**	**God's Eye**
Total Profit (All Matches)	$3,150	$19,937.5

Table 3 | Comparison of top 5 biggest "cash cows" from correct predictions

Model (XGB)		**God's Eye**	
Match	*Profit*	*Match*	*Profit*
CAM v. BRA	$1732.5	**ARG v. KSA**	$5880.0
JPN v. CRC	$1367.5	**CAM v. BRA**	$1732.5
AUS v. DEN	$1345.0	**JPN v. CRC**	$1367.5
JPN v. ESP	$1260.0	**AUS v. DEN**	$1345.0
KOR v. POR	$565.0	**GER v. JPN**	$1320.0

The betting framework from 5.1 placed bets across 20 matches; 14 had negative profits (i.e., predicted did not equal actual outcome) and 6 were correctly predicted. Of the top five cash cows from "God's Eye", the model correctly hit 3 of 5. Of note, the model correctly predicted "interesting" matches such as Cameroon v. Brazil. On the other hand, although not shown here, the model also struggled to predict some obvious matches such as Canada v. Morocco. Nonetheless, the positive profit suggests the betting framework places bets on matches where the expected returns are sufficiently high to warrant placing a bet on more unexpected outcomes, which did occur (e.g., Cameroon v. Brazil). However, the total profit pales in comparison to "God's Eye" but this is expected as perfect knowledge cannot be assumed in reality.

6.0 Knockout Stage Predictions

Using the same XGB model for the betting framework in 5.0, predictions for the knockout stages were also generated to determine an overall winner for the 2022 FIFA World Cup. At the time of writing, all 8 teams progressing to the Quarter Finals (QF) from the Round of 16 (RO16) were correctly identified. Of note, the model also correctly predicted that Morocco v. Spain went to penalties, with Morocco winning on penalties. Overall, France are predicted to be the world champions, Brazil as runner ups, Argentina third, and Portugal in fourth. The predictions for the knockout stages are shown in Figure 3 below.

Figure 3 | Predicted knockout stages for the 2022 FIFA World Cup

7.0 Conclusions

While the combination of the Machine Learning models in tandem with the betting framework has proven to have some predictive skill and benefit, numerous improvements could be made. For example, one could incorporate a “penalty term” in the objective function that penalizes the expected profit based on predicted winners, losers, and probabilities. Overall, predicting the outcomes of the World Cup is difficult as “obvious winners” may not always be the actual winner (such as Argentina v. Saudi Arabia). While applying an expected profit betting framework may help, the results shown in this report have demonstrated that it is extremely difficult to come close to correctly predicting the outcomes of every match.

8.0 Appendix A: "God's Eye" Betting Strategy

For "God's Eye," the probabilities of each group stage match *i* and outcome *j* are not required as the outcome of each match is "known." This "known" outcome also means the *j* dimension is not required.

Sets

- Matches: i = 1, ..., 48

Data

- Earnings per dollar on a group stage match *i* for the actual outcome: R_i

Decision Variables

- x_i: Betting amount on match *i*
- z_i: 1 if a bet is made on match *i*, 0 otherwise

Objective Function

$$max \sum_{i=1}^{48} (R_i x_i) - \sum_{i=1}^{48} x_i$$

Constraints

1) Bet cannot exceed 5% (1/20) of the budget on any match *i*

$$x_i \leq \frac{B}{20} z_i \qquad \forall i$$

2) Total bet cannot exceed available budget, B

$$\sum_{i=1}^{48} x_i \leq B$$

3) Risk spreading over at least *K* matches

$$\sum_{i=1}^{48} z_i \geq K$$

4) If $z_i = 1$, a betting amount x_i must be made

$$x_{ij} \geq z_{ij}$$

9.0 Appendix B: Group Member Contributions

Shannan focused on data processing, producing the models for predicting match outcomes, and gathering the geographic and demographic data via web scraping (weather, city populations)

Yen acquired international match data via web scraping, downloading datasets from kaggle. Yen also produced the betting framework and the initial optimization model formulation.

Yen and Shannan collaborated on fine-tuning the optimization formulation and producing the World Cup's group-stage, quarter-final, semi-final, and final stage predictions.

The report and slides were produced collaboratively, with each person writing about their main area of work and combining it.

10.0 References

FIFA (2018) *€136Bn Betting Turnover and no suspicious betting behavior at Russia 2018*. Available from: https://www.fifa.com/tournaments/mens/worldcup/2018russia/news/136bn-betting-turnover-and-no-suspicious-betting-behaviour-at-russia-2018 [Accessed: 22nd November 2022]

FIFA (2018) *More than half the world watched the record-breaking 2018 World Cup*. Available from: https://www.fifa.com/tournaments/mens/worldcup/2018russia/media-releases/more-than-half-the-world-watched-record-breaking-2018-world-cup [Accessed: 22nd November 2022]