To Bet or Not to Bet: Analyzing the Credibility of Fixed-odds Betting on Match Outcomes

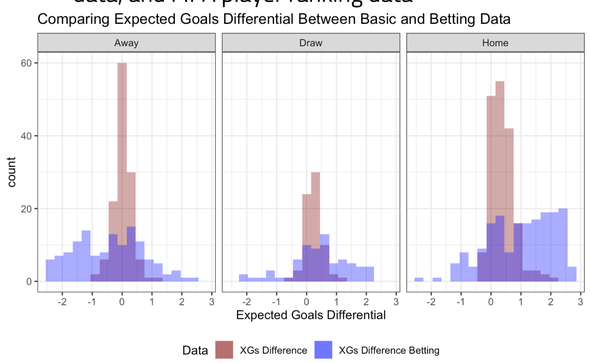
Maria Tsakalakos, Fungai Jani, and Tseegi Nyamdorj
Carnegie Mellon University

INTRODUCTION

The world of soccer is constantly growing, with millions of fans all over the globe. With the growing fanbase comes a growing betting community. Soccer accounts for 70% of all global sports bets placed. We focus specifically on the Premier League, as it is the most-watched soccer league. This paper aims to build a comprehensive predictive framework for forecasting match results in the Premier League season of 2018-2019. To build upon the already existing foundation of betting odds we enriched our predictive models by adding two key components: player evaluation metrics and the teams' recent form based on their performances over the last five matches.

DATA SOURCES / EDA

- Our data is comprised of game-to-game Premier League data in the 2018-19 season
- We use three data sets: Basic game data, betting data, and FIFA player ranking data



- The betting dataset does a better job of displaying the expected goal differential than the basic dataset.
  - This is because curves from the basic data set are normal distributions, whereas the curves from the betting data are skewed left or right, depending on whether the home or away team wins.
- By incorporating betting data in our models to predict outcomes of Premier League matches, we can achieve more precise predictions because of the skewness within the betting data

METHODOLOGY

Our base models is to predict the match outcomes through betting-odds. As the dependent variable (match outcome) is an ordered categorical variable, the models we could utilize were inherently limited to main three models: Generalized Additive Models (GAMs), Multinomial regressions, and Random forest algorithms. After assembling the base models, we added more independent parameters to increase the accuracy prediction of our models. In total, we used four different explanatory variables, namely the streak difference (the last five games' results of each team leading up to that specific match), expected goals (XGs) for the home team computed by the betting odds, XGs difference, and XGs difference based on defense and offense rating of home/away teams. One of the main objectives of our project is to incorporate team ratings based on player evaluation metrics to better the accuracy of models. We have decided to use a multiplicative regression of offense, defense ratings, and XGs.

$$min\Sigma_i^n[(y_{hi} - y_{ai}) - (g_{h_i} \cdot o_h \cdot d_h - g_{a_i} \cdot o_a \cdot d_a)]$$

After setting up the models, we trained the models on matches from the start of the season until February of the 2018/19 season as it encompasses ⅔ of the season and allows machine learning to best adapt to the available data, which then can be used to test out the remaining 2018/19 season matches.

RESULTS

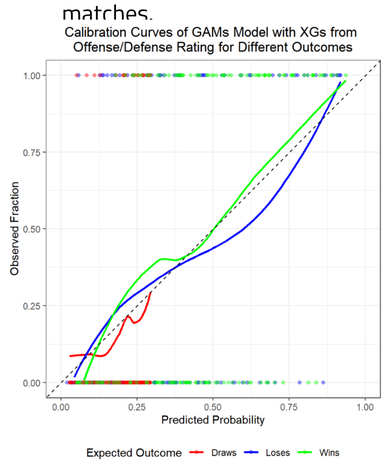
In order to best test out the accuracy of model predictions, we computed each models' brier score, an equivalence of the mean squared error as applied to predicted probabilities.

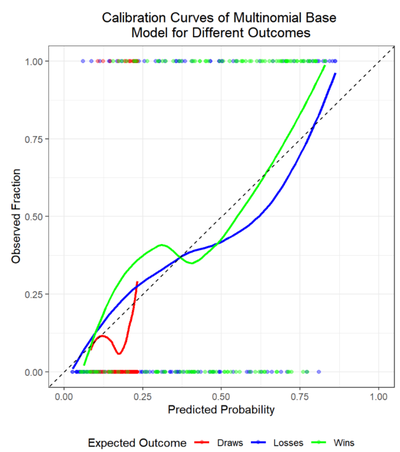
The best model (lowest brier score) is the GAMs with XGs based on defense and offense ratings on the baseline model. Although most brier scores range between 0.25 and 0.27, scores of random forest type have exceptionally high brier scores, averaging around 0.46. Thus, we can infer that the GAMs is the best applicable method, followed closely by the multinomial models.

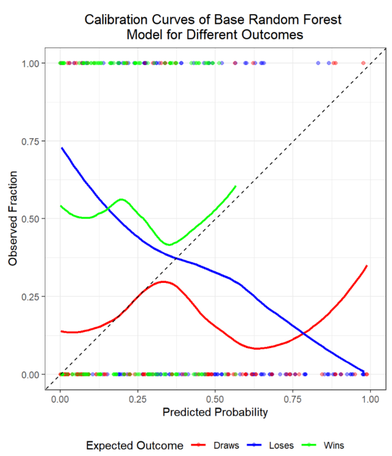
	Brier_Score	Type_of_Model
	<dbl>	<chr>
Pos GAMs	0.2552390	GAMs
multinomial betting pos	0.2565047	multinomial
base GAMs	0.2574400	GAMs
multinomial full	0.2583760	multinomial
rand forest pos	0.4395077	random forest
rand forest streak Xg	0.4444529	random forest
rand forest base	0.4993781	random forest

Moreover, in order to best represent the model accuracy, we decided to use calibration plots.

- The best model (GAMs) has its class lines predominantly closer to the dashed line of perfect calibration.
- Multinomial base models over-predicts match outcomes causing type II errors.
- The base random forest has both type I and type II errors, which means the model is unfavorable for predicting match outcomes.







CONCLUSION

In this project we had the goal of enhancing match outcome predictions in the 2018-2019 English Premier League season by using a variety of predictive models and including more variables. We aimed to outperform the standard betting probabilities and used this as our baseline model. Our results showed that adding player and team evaluation to the baseline model, considerably increased the accuracy of match outcome forecasts. Through the utilization of generalized additive models, random forest and multinomial models, we gained valuable insights into the strength and weaknesses of each approach.

Despite the promising results and contributions, there are limitations to this project namely: the spontaneity of sports events, particularly soccer matches. While we incorporated recent form and player evaluations, unpredictable factors such as injuries, weather and unexpected tactical adjustments during the game. These dynamic elements may limit the accuracy of our predictions. Furthermore, the data is specific to one English Premier League season, future seasons may show different trends and dynamics.

ACKNOWLEDGEMENTS

For their direction, inspiration and constant support, we are grateful to Dr. Konstantinos Pelechrinis, Dr. Ronald Yurko and Meg Ellingwood. Their assistance throughout this project has been immensely appreciated. We also extend our thanks to the entire Carnegie Mellon Sports Analytics Camp of 2023, the guest speakers and the Teaching Assistants. Finally, we are thankful to Carnegie Mellon University for funding this research camp and project.