Project Report

Predicting Football Match Outcomes Using Betting Odds and Elo Ratings

Arda Gökmen Yiğit – DSA 210 Term Project

Project Goal and Motivation

The goal of this project is to investigate how accurately football match outcomes can be predicted using pre-match betting odds and team Elo ratings. By combining statistical analysis and machine learning techniques, the project aims to:

- Examine the predictive power of betting markets.
- Assess whether Elo rating differences between teams improve prediction performance.
- Provide insights for data scientists and sports analysts regarding model selection and feature importance in outcome prediction.

Dataset and Preprocessing

The dataset includes official match results, pre-match betting odds, and Elo ratings for both home and away teams. The following features were used:

- ProbH / ProbD / ProbA: Implied probabilities for Home win, Draw, Away win, derived from odds.
- HomeElo / AwayElo: Elo scores representing team strength.
- EloDiff: Difference between HomeElo and AwayElo.

Preprocessing Steps:
- Conversion of date fields into datetime format (if applicable).
- Cleaning missing values and standardizing odds format.
- Calculating implied probabilities from decimal odds.
- Creating EloDiff = HomeElo - AwayElo to capture relative team strength.
- Normalizing the data for uniform scale was considered but not necessary as the models used are tree-based and robust to scale.

Statistical Analysis and Hypothesis Testing

A binomial hypothesis test was conducted to assess the accuracy of implied probabilities.
Null hypothesis: Betting odds do not significantly predict match outcomes.
After conducting the analysis:

- Observed win rate for the highest implied probability outcome aligned significantly with actual match results.
- P-value was found to be less than 0.05, leading to rejection of the null hypothesis, supporting that betting odds contain significant predictive power.

Additionally, exploratory data analysis showed that:
- Home teams generally have higher implied win probabilities.
- Elo rating differences correlate with betting odds, confirming market awareness of team strength.

Machine Learning Model and Evaluation

A Random Forest Classifier was used to predict the categorical outcome (FTR: Home/Draw/Away) using the following input features:

- ProbH, ProbD, ProbA, EloDiff, HomeElo, AwayElo

Training and Testing:
- Dataset was split 80/20 into training and test sets.
- Accuracy, classification report, and confusion matrix were used for evaluation.
- The model achieved an accuracy of approximately 44%, which is reasonable for a 3-class prediction problem.
- Feature importance analysis revealed:
  - ProbH and ProbA were the most influential features.
  - EloDiff also had a meaningful contribution to prediction performance.

Key Results

- Betting odds alone provide a strong baseline for match prediction.
- Adding Elo ratings increased model accuracy and offered an interpretable improvement in prediction logic.
- Random Forest performed well with minimal hyperparameter tuning, and its feature importance metric provided clear insights.

Limitations and Future Work

Limitations:
- Model performance may be limited by data quality (e.g., missing matches, outdated Elo).
- Draw predictions remain challenging due to the nature of football and market uncertainties.

Future Enhancements:
- Add more features: weather, injuries, recent team form.
- Try ensemble models such as XGBoost or Voting Classifier.
- Deploy the model with live updating odds using real-time APIs.

Conclusion

This project successfully demonstrates that a combination of betting odds and Elo ratings can be used to build a reasonably accurate predictive model for football match outcomes. Through statistical analysis and machine learning, we validated the predictive value of both market data and team strength metrics.
The model is a solid foundation for more advanced sports analytics and future real-time prediction systems.