A Machine Learning Approach for Football Match Prediction Using Comprehensive Feature Engineering

Preliminary Report

Berkay Bakisoglu
Department of Computer Engineering
Ege University
Izmir, Turkey
Email: 91230000563@ogrenci.ege.edu.tr

***Abstract*—In this paper, we explore the development of a machine learning-based match prediction system for football. Our primary goal is to create an accurate prediction system that can effectively forecast match outcomes across different leagues. We propose a comprehensive approach that combines feature engineering with various machine learning models, processing extensive historical match data (2010-2024) with particular attention to team performance patterns and historical statistics. While this is a preliminary report, our initial analysis suggests promising directions not only for accurate match prediction but also for potential applications in betting markets. We discuss our methodology, current progress, and planned experiments for validating our approach.**

I. Introduction

Football match prediction presents a complex challenge for machine learning applications. Despite the abundance of historical match data and statistics, accurately predicting match outcomes remains difficult due to the numerous variables involved and the dynamic nature of team performance. Our research began with the goal of developing a reliable match prediction system that can effectively process historical data and identify patterns in team performance. Beyond the primary goal of accurate prediction, we also aim to explore whether such a system could be effectively applied in betting scenarios.

A. Problem Statement

Through our initial research, we identified several key challenges:

- Extracting meaningful patterns from complex match statistics and historical data
- Developing models that can adapt to changing team performance and form
- Creating effective evaluation methods that consider both prediction accuracy and consistency
- Exploring the practical applications of accurate predictions in betting contexts

B. Research Objectives

Our research has two main goals:

1) Primary Objectives:

- Development of an accurate match prediction system for multiple leagues
- Implementation of comprehensive feature engineering focusing on team performance metrics
- Creation of robust evaluation metrics for prediction accuracy
- Analysis of different machine learning approaches for match prediction

2) Secondary Objectives:

- Evaluation of prediction system's applicability to betting scenarios
- Analysis of prediction confidence in relation to betting decisions
- Investigation of system performance in different betting markets

II. Previous Works

Sports betting prediction using machine learning has gained significant attention in recent years. Terawong & Cliff (2024) demonstrated the effectiveness of XGBoost in learning profitable betting strategies through an agent-based model of a sports betting exchange. Their work showed that machine learning models could learn strategies that outperform traditional betting approaches, achieving an overall accuracy of 88%.

Bunker & Thabtah (2017) proposed a structured framework for sports result prediction (SRP-CRISP-DM) that emphasizes the importance of proper data preprocessing and feature engineering. Their framework distinguishes between match-related and external features, and advocates for preserving temporal order in model evaluation.

These works highlight two key aspects in sports betting prediction: the importance of sophisticated machine learning approaches and the need for proper data handling and evaluation methodologies. Our work builds upon these foundations while introducing several novel elements:

- A hierarchical prediction system that leverages correlations between different betting markets

- Enhanced feature engineering incorporating both historical statistics and market-derived probabilities
- A comprehensive evaluation framework that considers both prediction accuracy and betting profitability

III. Methodology

This section presents our approach to football match prediction, which combines standardized feature engineering with three distinct model architectures. The methodology is structured to enable fair comparison between different prediction approaches while maintaining architectural innovation.

A. Feature Engineering

The system employs a standardized set of features across all predictors, ensuring fair model comparison. Features are organized into:

1) Base Features: Team performance metrics including goals scored/conceded, clean sheets, and win rates. Form features capture recent performance through exponentially weighted averages. League position features track team standings and points.

2) Market Features: Market-derived features include implied probabilities and value indicators for match outcomes (Home/Draw/Away) and total goals (Over/Under 2.5). Market confidence metrics and overround calculations assess betting efficiency.

B. Prediction Systems

Three distinct architectural approaches share the standardized feature set:

1) Unified Predictor: Combines Random Forest and LightGBM models for independent predictions across markets. Handles both match outcomes and over/under predictions through separate classifiers.

2) Hierarchical Predictor: Sequential LightGBM models predict auxiliary markets first (cards, corners), enhancing match outcome predictions. Maintains separate models for over/under predictions.

3) Weighted XGBoost Predictor: Employs market-odds weighted XGBoost models. Handles match outcomes and over/under markets with specialized sample weights based on implied probabilities.

C. Evaluation Framework

Performance assessment includes:

- Classification metrics (accuracy, precision, recall) for match outcomes and over/under predictions
- Return on Investment (ROI) calculations for each market
- Value bet identification based on predicted probabilities versus market odds
- Market-specific analysis comparing performance across different betting types

D. Model Comparison Framework

To ensure scientific rigor in comparing these architectures:

- All models use identical feature sets
- Time-series based validation preserves temporal order
- Performance metrics include accuracy, ROI, and market-specific measures
- Analysis of model strengths across different leagues and seasons

This methodology enables us to evaluate the effectiveness of different architectural approaches while maintaining consistency in feature engineering and evaluation metrics.

IV. Experimental Studies

A. Dataset and Implementation Details

1) Dataset Characteristics: Our dataset comprises historical football match data from major European leagues:

- **Data Source**: Historical match data and betting odds from 2010 to 2024, football-data.co.uk
- **Dataset Size**: 15 seasons of data across multiple leagues
- **Features**: Over 30 features per match including:
  - Match statistics (goals, corners, cards)
  - Team performance metrics
  - Historical head-to-head records
  - Betting odds from Bet365
- **File Format**: Mix of .xls and .xlsx files, organized by season and league

2) Implementation Environment: The system was implemented using the following technologies:

- **Programming Language**: Python 3.10+
- **Key Libraries**:
  - scikit-learn (for Random Forest models)
  - LightGBM (for gradient boosting)
  - pandas (for data manipulation)
  - numpy (for numerical operations)
  - matplotlib and seaborn (for visualization)
- **Development Environment**: Pycharm

B. Training Methodology

1) Data Preprocessing: Our preprocessing pipeline includes:

- Temporal alignment of match data
- Feature scaling using StandardScaler
- Handling of missing values through predefined rules
- Validation of data completeness and consistency

2) Cross-Validation Strategy: We implemented a time-series based cross-validation approach:

- **Training Mode**:
  - Minimum 3 seasons required for validation
  - Sliding window approach for season selection
  - Sequential split to maintain temporal order
- **Test Mode**:
  - 80% training, 20% testing split
  - Configurable test size parameter
  - Rapid prototyping capabilities

C. Performance Metrics

We evaluate our models using multiple metrics:

- **Prediction Accuracy Metrics**:
  - Accuracy: Overall prediction accuracy for match outcomes
  - Precision: Accuracy of positive predictions
  - Recall: Ability to detect positive cases
  - F1-score: Harmonic mean of precision and recall
- **Betting Performance Metrics**:
  - Return on Investment (ROI): Percentage return on placed bets
  - Profit/Loss: Absolute monetary performance
  - Strike Rate: Percentage of successful bets
  - Value Betting Analysis: Comparison of predicted vs. market probabilities
- **Market-Specific Metrics**:
  - Match Results: Classification metrics with ROI per outcome
  - Corners/Cards: RMSE and MAE for count predictions
  - Market Efficiency: Analysis of odds-implied vs. predicted probabilities

These metrics provide a comprehensive view of both predictive accuracy and betting effectiveness, enabling evaluation of models from both statistical and practical perspectives.

D. Analysis Tools

Our evaluation framework includes:

- **Seasonal Progression Analysis**:
  - Tracking prediction accuracy over time
  - Analyzing model adaptation to season changes
- **League-Specific Analysis**:
  - Performance comparison across leagues
  - League-specific feature importance
- **Feature Importance Analysis**:
  - Ranking of most influential features
  - Market-specific feature analysis

V. Experimental Results

A. Preliminary Setup

Our experimental framework addresses both prediction accuracy and betting applications:

1) Planned Experiments: We aim to answer several key questions:

- How do different models perform in predicting match outcomes?
- Which features are most important for accurate prediction?
- How does prediction accuracy vary across different leagues?
- What is the impact of historical data window size on prediction accuracy?
- How well do accurate predictions translate to betting success?
- Which types of predictions offer the best betting opportunities?

2) Initial Data Analysis: Our preliminary investigation has focused on understanding our data:

- Analyzing match outcome distributions across leagues
- Understanding team performance patterns
- Examining Bet365 odds patterns and distributions
- Evaluating feature correlations with match outcomes

VI. Conclusion

This preliminary report represents our first steps toward developing an effective machine learning approach to football match prediction, with potential applications in betting markets. While we're still early in our research, our initial work has revealed both promising directions and significant challenges.

A. Current Progress

We've established some fundamental building blocks:

- A robust approach to processing historical match data
- A framework for extracting meaningful performance features
- Initial prediction model architectures
- A comprehensive evaluation methodology

B. Future Work

Our next phase of research will focus on several key areas:

1) Model Development:

- Testing and comparing different model architectures
- Implementing ensemble methods for improved prediction accuracy
- Developing more sophisticated confidence estimation techniques

2) Validation and Testing:

- Testing performance using historical betting odds

3) Betting Applications:

- Analyzing prediction accuracy in relation to betting profitability

4) Research Extensions:

- Expanding our literature review and theoretical foundation
- Analyzing prediction performance across different leagues and seasons
- Investigating market-specific prediction strategies
- Developing real-time prediction capabilities