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

Our approach introduces a novel hierarchical prediction system for football betting markets. The system consists of four main components:

A. Data Processing

The data processing pipeline handles multiple seasons of football match data from various European leagues. Our dataset contains comprehensive match statistics and betting odds, including:

- **Match Statistics**: Full-time (FT) and half-time (HT) scores, shots on target (HST/AST), fouls (HF/AF), corners (HC/AC), and cards (HY/AY/HR/AR)
- **Betting Odds**: Pre-match odds from multiple bookmakers (Bet365, BetWin, Betfair, etc.) for:
  - Match results (Home/Draw/Away)
  - Over/Under 2.5 goals
  - Asian Handicap markets
  - Corner markets

B. Feature Engineering

Our feature engineering pipeline creates sophisticated predictive features from the raw data:

- **Team Performance Features**:
  - Recent form with exponential decay weights
  - Rolling averages for goals scored/conceded
  - Team-specific metrics (clean sheets, scoring patterns)
  - Days since last game for recovery analysis
- **Market-Derived Features**:
  - Implied probabilities from betting odds
  - Market overround calculations
  - Value betting indicators
  - Favorite/underdog identification
- **League Position Features**:
  - Dynamic league standings
  - Points and position differences
  - Goal difference rankings
  - Form-based league performance
- **Match-Specific Features**:
  - Head-to-head statistics
  - Home/Away performance metrics
  - Referee influence indicators
  - Derby match identification
- **Advanced Statistical Features**:
  - Standard deviation of performance metrics
  - Weighted historical encounters
  - Seasonal progression indicators
  - League-specific normalization

The feature engineering process includes careful handling of temporal aspects to prevent data leakage, with all features calculated using only historical data available before each match. Feature importance analysis guides the selection of most relevant predictors for each market.

C. Hierarchical Prediction System

The core of our system is a novel hierarchical approach that combines multiple prediction models:

- **Base Predictors**: Specialized models for corners and cards using LightGBM and Random Forest
- **Enhanced Match Predictor**: A final classifier that incorporates predictions from base models
- **Confidence Estimation**: Uncertainty quantification using model-specific techniques

D. Evaluation Framework

Our evaluation framework provides comprehensive assessment through:

- **Time-Series Validation**: Proper handling of temporal dependencies
- **Market-Specific Metrics**: Tailored evaluation metrics for each prediction target
- **League-Specific Analysis**: Performance breakdown by league and season
- **Visualization Tools**: Advanced plotting capabilities for result analysis

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. Model Parameters

1) Random Forest Configuration: For classification tasks (match results):

- n_estimators: 200
- max_depth: 15
- min_samples_split: 10
- min_samples_leaf: 5
- class_weight: 'balanced'

2) LightGBM Configuration: For regression tasks (corner-card prediction):

- n_estimators: 500
- learning_rate: 0.01
- num_leaves: 31
- max_depth: 8
- min_child_samples: 20

D. Performance Metrics

We evaluate our models using multiple metrics:

- **Classification Tasks**:
  - Accuracy
  - Precision
  - Recall
  - F1-score
- **Regression Tasks**:
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - Mean Absolute Error (MAE)
  - R-squared (R2)

E. 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

B. Next Steps

Our immediate plans include:

- Testing different model architectures
- Analyzing feature importance across leagues
- Validating prediction accuracy with historical data
- Comparing our predictions with Bet365 implied probabilities
- Evaluating prediction confidence thresholds
- Testing performance using historical Bet365 odds

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 will focus on:

- Implementing and testing various prediction models
- Conducting extensive validation experiments
- Expanding our literature review
- Analyzing prediction performance across different leagues