From Opening Day to Season's End: The Evolution of Predictive Model Performance in European Football Match Outcome Prediction

Jannik Herrlich

Dissertation written under the supervision of professor Domenico Fabrizi

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, 05.01.2026.

Abstract

This dissertation analyses how the predictive performance of football match outcome models evolves over the course of a season as additional information becomes available. While most existing studies evaluate predictive models using full-season data under the assumption of static conditions, this study explicitly focuses on temporal performance dynamics and the point at which predictions become stable and reliable. Match-level data from six professional European football leagues across two competitive tiers are examined, including a structurally regular season (2023/2024) and a disrupted season affected by the COVID-19 pandemic (2019/2020). Four modeling approaches are evaluated: Multinomial Logistic Regression (MLR) as a linear baseline, Random Forest (RF) and Gradient Boosting (GB) as non-linear tree-ensemble models, and a Multi-Layer Perceptron (MLP) representing a deep learning approach. Model performance is assessed using time-based splits within each season, allowing predictive accuracy, discrimination and calibration to be analysed across successive phases of the season. To ensure methodological consistency and comparability, all models are trained using default hyperparameter settings and identical preprocessing pipelines. Performance is evaluated using Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC) and Logarithmic Loss, with particular emphasis on probabilistic calibration and prediction stability. The results indicate that predictive performance improves over the season, with ensemble-based models converging more rapidly toward stable performance than linear models. Differences across leagues and competitive tiers are most pronounced early in the season, highlighting the importance of incorporating temporal considerations into football outcome prediction.

Title: From Opening Day to Season's End: The Evolution of Predictive Model Performance in European Football Match Outcome Prediction

Author: Jannik Herrlich

Keywords: Football Match Outcome Prediction; Predictive Analytics; Temporal Performance Dynamics

Resumo

Esta dissertação analisa como o desempenho preditivo de modelos de previsão de resultados de partidas de futebol evolui ao longo de uma temporada à medida que novas informações se tornam disponíveis. Enquanto a literatura existente normalmente avalia modelos preditivos utilizando dados de temporadas completas e assumindo condições estáticas, este estudo concentra-se explicitamente na dinâmica temporal do desempenho dos modelos e no momento em que as previsões se tornam estáveis e confiáveis. A análise utiliza dados em nível de partida de seis ligas profissionais europeias, abrangendo dois níveis competitivos, incluindo uma temporada estruturalmente regular (2023/2024) e uma temporada afetada por irregularidades decorrentes da pandemia de COVID-19 (2019/2020). Quatro abordagens de modelagem são avaliadas: Regressão Logística Multinomial como modelo linear de referência, Random Forest e Gradient Boosting como métodos não lineares baseados em ensembles de árvores, e uma Rede Neural do tipo Multi-Layer Perceptron representando uma abordagem de aprendizado profundo. O desempenho dos modelos é avaliado por meio de divisões temporais ao longo da temporada, permitindo a análise da acurácia, capacidade discriminativa e calibração das previsões em diferentes fases do campeonato. Para garantir consistência metodológica, todos os modelos utilizam configurações padrão de hiperparâmetros e pipelines de pré-processamento idênticos. A avaliação do desempenho é realizada com as métricas Accuracy, AUC e Logarithmic Loss, com ênfase na calibração probabilística e na estabilidade das previsões. Os resultados mostram que o desempenho melhora ao longo da temporada, com modelos baseados em ensembles atingindo estabilidade mais rapidamente, especialmente quando comparados a modelos lineares.

Título: Do Início da Época ao Final da Temporada: A Evolução do Desempenho de Modelos Preditivos na Previsão de Resultados de Jogos de Futebol Europeu

Autor: Jannik Herrlich

Palavras-chave: Previsão de Resultados de Jogos de Futebol; Análise Preditiva; Dinâmica Temporal do Desempenho

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Domenico Fabrizi, for his guidance, constructive feedback and continuous support throughout the development of this dissertation. His insights and critical perspective were essential in shaping the research design, refining the analysis and improving the overall quality of this work. I would also like to thank my group members for their constructive feedback during the research process.

I am especially grateful to Prof. Filipa Reis, as well as to all professors of the Business Analytics faculty, for providing a stimulating academic environment and equipping me with the analytical and methodological skills necessary to conduct this research.

Furthermore, I would like to thank my family and Irem Canbolat for their encouragement, patience and support throughout the writing process. Their understanding and motivation were invaluable, particularly during the more demanding phases of this project.

Finally, I would like to acknowledge all those who contributed indirectly to this work, whether through discussions, feedback or technical assistance. Their contributions are greatly appreciated.

Contents

1 Introduction

1.1 Background and Significance

Football has long been the world's leading sport, capturing the attention of billions across diverse regions and cultures.

Historically associated with working-class culture, football has transformed into one of the most commercially powerful industries in the world. Nowhere is this commercialization more visible than in European football, which has become the focal point of global investment, media coverage and competitive market dynamics. European football generates enormous financial value, increasing the pressure on clubs to achieve consistent success. As a result, data analytics is gaining importance in strategic decision-making within clubs. At the same time, betting markets, which are highly dependent on accurate forecasts, have shown growing interest in advanced football analytics too. With thousands of matches played each season, the ability to forecast match outcomes has become central to tactical planning, market pricing and analytical decision-making. Recent advances in data analytics and machine learning have made predictive modeling increasingly accessible.

However, the temporal dynamics of predictive performance throughout a season remain poorly understood. Understanding how model performance evolve during the season is essential, since many existing approaches neglect intra-seasonal changes in team strength, form and context. Ignoring these dynamics leads to less interpretable and potentially misleading predictions. Teams change over the season in form of their performance levels, player injuries or transfers. Early-season predictions are inherently uncertain due to limited data availability and accumulate only gradually as the season progresses.

This changing information base affects model performance over time. Understanding these dynamics is crucial for practical applications, whether in football analytics or in betting-related decision-making.

1.2 Motivation

The increasing use of quantitative metrics and key performance indicators in football analysis highlights the growing relevance of predictive analytics in understanding match outcomes. While most existing studies evaluate predictive models using full-season data under the assumption

of static conditions, this dissertation focuses on how predictive performance evolves over the course of a season as new information becomes available.

By analysing when and how models begin to produce stable and reliable forecasts, this research contributes to a deeper understanding of match outcome prediction models and their practical applicability. The findings are relevant for stakeholders such as sports analysts and bettors, who rely on probabilistic forecasts for decision-making under uncertainty.

1.3 Research Questions

To obtain meaningful insights into the dynamics of predictive modeling in football, three research questions are formulated.

The first and most central question asks: 'How do predictive performance metrics of established models evolve from the beginning to the end of the season as more matches are played?'. This question aims to capture the core phenomenon investigated in this dissertation, how a model's predictive ability changes as the season progresses and additional match information becomes available. Understanding this evolution provides the foundation for all subsequent analyses and enables a systematic examination of performance trends from the opening matchday to the end of the season.

The second research question asks: 'Are there systematic differences in time-to-stable prediction across league tiers and across seasons?' This question examines whether predictive performance stabilizes at different points in the season depending on the competitive level and season of the league. By comparing league tiers and different seasons, the analysis examines whether structural characteristics, such as differences in team quality, competitive balance or variability in match outcomes, affect how quickly predictive models reach consistent and reliable performance.

The final research question asks: 'Do simple Multinomial Logistic Regression (MLR), Random Forest (RF), Gradient Boosting (GB) and Multi-Layer Perceptron (MLP) model adapt differently to seasonal information growth?'. This question investigates whether model complexity influences how predictive performance stabilize throughout the season. By comparing a relatively simple and interpretable MLR model, well suited for multi-class outcomes such as win-draw-loss, with more complex approaches, including the non-linear tree-ensemble models (RF and GB) and a MLP as a deep learning model, this analysis investigates whether these four modeling approaches respond differently to the gradual accumulation of match data. Specifically,

it explores which model type benefits earlier or more strongly from increasing information and how their predictive performance stabilize over time.

1.4 Research Hypotheses

Based on research questions and theoretical considerations surrounding predictive modeling in football, several hypotheses are developed. These hypotheses aim to formalize expectations about how predictive performances evolve throughout the season, how league structure influences performance stability and how different modeling approaches adapt to increasing information.

H1: Seasonal Improvement Hypothesis (RQ1)

As more match data become available throughout the season, predictive models are expected to improve in both discrimination and calibration. Early-season predictions suffer from limited information and unstable team signals, whereas mid- and late-season predictions benefit from more accurate representations of team strength, form and contextual dynamics.

H1a: Performance Metrics improve with increasing Data

Discrimination metrics (Accuracy, AUC) and calibration metrics (LogLoss) improve significantly from early-season (T1) to late-season (T4), reflecting decreasing model uncertainty and more stable learning conditions.

H2: Tier stability Hypothesis (RQ2)

Differences in competitive structure between first-tier and second-tier leagues are expected to affect how quickly predictive models reach stable performance.

H2a: First-tier leagues stabilise earlier

Predictive performance in first-tier leagues reaches stability earlier in the season compared to second-tier leagues. Second-tier leagues exhibit greater early- and mid-season volatility in both discrimination and calibration metrics, due to higher competitive randomness and greater team turnover.

H3: Model complexity adaptation Hypothesis (RQ3)

Different model families adapt differently to the accumulation of seasonal information due to their varying complexity and capacity to learn nonlinear relationships.

H3a: Tree-ensemble models show stronger performance gains

Tree-ensemble models (RF, GB) are outperforming the MLR and MLP models in early- and mid-season due to their ability to capture nonlinear interactions in team-strength and form-related variables. As additional match data become available, performance differences between models will diminish or converge, as all models benefit from increased sample size and more stable patterns.

1.5 Research Objectives and Scope

After outlining the background of the dissertation topic, presenting the motivation and clarifying the main research questions and its hypotheses, it is essential to define the specific objectives and scope of this scientific work.

1.5.1 Objective

The primary objective of this dissertation is to analyse how the predictive performance of football match outcome models evolves over the course of a season as additional match data become available. The study focuses on both discrimination and calibration performance and examines when predictive models begin to produce stable and reliable forecasts.

A further objective is to compare predictive performance dynamics across league tiers, seasons and model families in order to identify systematic differences in learning behaviour and time-to-stability.

1.5.2 Scope

This dissertation addresses a multi-class classification problem, predicting match outcomes as home win, draw or away win. The empirical analysis is conducted using match-level data from

the 2019/2020 and 2023/2024 seasons of six European football leagues across two competitive tiers.

Four modeling approaches are evaluated: MLR, RF, GB and a MLP. Model performance is assessed using Accuracy, Area under the ROC Curve (AUC) and Logarithmic Loss (LogLoss).

The analysis is restricted to structured match-level information and pre-match betting odds. Player-level data, tracking data and in-game event data are not considered.

2 Literature review

Football outcome prediction has long attracted the interest of academia, bookmakers and professional clubs due to its significant financial and tactical implications. The proliferation of machine learning techniques and enriched datasets over the last two decades has produced an expanding literature on statistical match forecasting (Tax & Joustra, 2015; Baboota & Kaur, 2018).

Researchers have explored a range of models, from early Poisson-based approaches (Dixon & Coles, 1997; Baio & Blangiardo, 2010) to modern ensemble learning and deep neural networks (Wong et al., 2025) for predicting outcomes such as home win, draw or away win. While many studies focus on maximizing predictive accuracy, there is growing interest in how models behave over time, especially as team form, strength and contextual variables evolve throughout the season (Macrì et al., 2023).

Despite the already wide breadth of existing work, relatively few studies examine how predictive performance metrics like Accuracy, AUC or calibration (LogLoss) shift from the beginning to the end of the season. This literature review synthesizes key developments in football match outcome modeling by focusing particularly on model dynamics, temporal stability and comparative performance between classic logistic regression, advanced machine learning models such as random forest forests and deep learning models (Bunker et al., 2023; Mills et al., 2024).

2.1 Classical Statistical Models in Football Predictions

The first steps into the field of football prediction were based on modified Poisson and bivariate goal distributions (Maher, 1982; Dixon & Coles, 1997; Macrì-Demarti et al., 2023). These models were designed to better capture the distribution of low-scoring matches and draws,

which are common in football. Unlike more recent models focused on predicting three-way outcomes (home win, draw, away win), these early approaches aimed to explain the underlying number of goals by each team (Macrì-Demarti et al., 2023). They treated goals as count variables and modeled them as a function of team-specific parameters, such as attacking and defensive strengths, to estimate the expected goal output for each side.

Martì-Demarti et al. developed an extension of this classical Poisson framework called Bayesian weighted discrete-time dynamic model. Their developed extension captures evolving team strengths over time by using a dynamic structure to address the non-stationarity of team quality throughout the season (Macrì-Demarti et al., 2023).

While Poisson models focus on goal scoring as a statistical process, they struggle to reflect dynamic changes in team strength. This led to the development of rating-based systems, which explicitly track evolving performance over time.

2.2 Rating-Based Models and Team Strength Evolution

In the next step in the field of football match prediction, rating-based systems were introduced, most notably those based on Elo ratings, which were originally developed to assess the relative strength of chess players (Elo, 1978). It is important to note that rating-based systems do not function as outcome prediction models. Rather, they estimate underlying team strengths through structured update rules and these ratings are then used as covariates or inputs in models that perform actual match-outcome prediction.

In football, Elo ratings were adapted to serve as dynamic covariates in match outcome prediction models, reflecting how the underlying team strength evolves throughout a season (Hvattum & Arntzen, 2010; Constantinou & Fenton, 2012). The Elo system update a team's rating after each match based on the result and the relative strength of the opponent, making it particularly suited for capturing form trends and performance shifts across time.

Another important development in football match outcome explanation is the pi-rating system, which estimates a team's strength based on relative discrepancies in match outcomes between opponents (Constantinou & Fenton, 2012). Unlike Elo, which adjusts ratings through win/loss outcomes, pi-ratings incorporate both scorelines and contextual performances differences to provide a more nuanced measure of relative superiority between teams. This approach offers an alternative way to dynamically capture team form and strength, making it a valuable input for

match outcome forecasting models.

In addition to these rating models, another approach emphasizes the importance of player-level features in forecasting football outcomes. Huang & Zhang (2023) explored the use of player statistics extracted from the EA Spots video game series to construct team-level profiles based on attributes such as passing accuracy, defensive strength and physicality. The model captured dynamic changes in team composition and quality over time through aggregating these metrics for starting line-ups. The results of implementing these player statistics demonstrated that incorporating such feature-based indicators significantly improved match outcome prediction accuracy, particularly if combined with contextual factors like betting odds. This prediction approach illustrates how data-driven representations of team strength can complement or even outperform traditional rating-based systems.

These explanatory models marked the baseline in the field of football outcome prediction. Although rating systems introduced dynamism, they still relied on hand-crafted rules. The rise of machine learning (ML) enabled data-driven modeling of complex, nonlinear patterns, leading to a new generation of predictive models.

2.3 Machine Learning Models for Match Outcome Prediction

Following the initial use of statistical models and rating systems for match outcome prediction, the field shifted towards ML. The use of ML models for this purpose dates back to the mid-2010s (Wong et al., 2025). The adoption of ML in sports analytics has led to significant advancements, enabling new approaches to analyzing and forecasting match outcomes based on historical data.

Existing research on match prediction focused on using various ML techniques to predict football match results based on team-level statistics, such as historical match results and goal differences (Al-Bustami & Ghazal, 2025). These studies typically used models such as logistic regression, decision trees and neural network models (Bunker et al., 2023; Al-Bustami & Ghazal, 2025; Atta Mills et al., 2024). Statistical models have been integrated with ML techniques, such as RF classifiers with Poisson-based rankings, to develop hybrid prediction approaches (Bunker et al., 2023). Studies such as those by Atta Mills et al. (2024), Al-Bustami & Ghazal (2025) and Groll et al. (2019) also used XGBoost models to improve the learning process of the models.

In addition to ML models, an increasing number of studies explore deep learning approaches (DL). Studies such as those by (Bunker et al. (2023)) and (Atta Mills et al. (2024)) state

that neural networks and hybrid DL models can capture complex nonlinear relationships within match data. While DL models such as LSTM leverage temporal dependencies in team performance, feed-forward networks learn data patterns from historical match data. The results of the mentioned studies showed that DL models are on a par with and often superior to ML models. To improve the performance of the model predictions, many studies, such as those by Baboota & Kaur (2018), Tax & Joustra (2016), Wong et al. (2025) and Atta Mills et al. (2024), have incorporated contextual variables such as rolling team form, win margin rates, attack/defence strength and betting odds. These additional variables are created through feature engineering, on which the ML and DL models rely heavily.

2.4 Model Evaluation Metrics and Calibration

A wide range of evaluation metrics has been applied in football match prediction research, including Accuracy, Precision, Recall, F1-score, AUC, LogLoss and the Brier score (Atta Mills et al., 2024; Bunker et al., 2023; Choi et al., 2023; Macrì-Demarti et al., 2023). These metrics vary in what they measure, some focus on classification performance, others on probabilistic accuracy and their suitability depends on the structure and purpose of the prediction task. Because football outcome prediction is a multi-class classification problem (home win, draw, away win) with inherent class imbalance, particularly in the draw category, selecting the appropriate evaluation metrics is essential (Atta Mills et al., 2024; Choi et al., 2023). In this dissertation, the focus lies on the metrics most appropriate for evaluating probabilistic multi-class predictions: Accuracy, LogLoss and AUC.

Accuracy provides the proportion of correct predictions relative to all predictions and is widely used to summarise overall classification performance (Atta Mills et al., 2024; Wong et al., 2025). However, it does not account for uncertainty in probability estimates and can mask class imbalance.

The LogLoss offers a more informative assessment for probabilistic models by penalising incorrect but confident predictions. The penalty increases exponentially as predicted probabilities diverge from the true label. A perfect LogLoss equals zero (Wheatcroft & Sienkiewicz, 2021; Bunker et al., 2023).

To evaluate discriminative performance, the AUC metric is used. The AUC score measures a model's ability to distinguish between classes by quantifying the trade-off between true positive

and false positive rates. In multi-class settings, the one-vs-rest approach is commonly applied (Baboota & Kaur, 2018).

Finally, calibration is crucial in football forecasting because probability estimates, not only class labels, are relevant for decision-making in contexts such as betting and risk assessment (Wheatcroft & Sienkiewicz, 2021; Dixon & Coles, 1997). Many ML models, including RF, do not inherently produce well-calibrated probabilities (Platt, 1999). For example, RF aggregate class votes from decision trees, which represent frequency rather than true probability. Therefore, post-hoc calibration techniques are often required to ensure reliable probabilistic forecasts (Wheatcroft & Sienkiewicz, 2021).

2.5 Temporal Evolution of Predictive Performance

The core of this dissertation concerns the temporal evolution of predictive performance, referring to how performance metrics such as Accuracy, calibration and discriminative ability change over time as training data accumulates. In the context of football match outcome prediction, model behaviour is strongly influenced by the amount of available match data: as more observations are incorporated, prediction variance decreases and the generalisation ability of the models improves (Koopman & Lit, 2017). In contrast, early-season predictions suffer from high noise, unstable team form and incomplete information. These patterns only stabilise gradually as the season progresses.

Existing football forecasting research does not systematically address the temporal evolution of predictive performance. Most studies evaluate models on full-season or multi-season datasets, thereby overlooking how performance develops within the season. Common works (Atta Mills et al., 2024; Choi et al., 2023; Koopman & Lit, 2017) assess overall predictive accuracy or calibration and acknowledge early-season volatility, but they do not quantify it.

Furthermore, no existing research develops or measures temporal stability thresholds, nor do current studies conduct systematic comparisons of temporal performance across different leagues or across multiple seasons.

2.6 Gaps and Justification for this Study

While a substantial body of research has examined the prediction of football match outcomes using statistical, rating-based and machine learning models, the question of how the predictive

performance of these models evolves over the course of a season has received very limited attention.

Existing studies typically evaluate model Accuracy, AUC-score or probabilistic scoring rules at the aggregate level, either across entire seasons or pooled over multiple seasons (Atta Mills et al., 2024; Wong et al., 2025; Al-Bustami & Ghazal, 2025). These approaches assess how well models can classify home wins, draws and away wins overall, but none of them analyse how prediction quality changes as more match data becomes available throughout a season. Even studies using dynamical statistical models or rating systems, such as Elo-based approaches (Hvattum & Arntzen, 2010), Bayesian dynamic Poisson models (Macrì-Demarti et al., 2023) and player-based team strength profiles (Huang & Zhang, 2023), do not investigate the temporal evolution of predictive performances in depth. Instead of measuring them empirically, they assume the temporal evolution theoretically. The only partial exception is the scientific paper by Tax & Joustra (2016), in which the authors developed a retrodictive framework for updating model parameters after each match.

However, this study does not analyse or track the progression of predictive performance over time, nor does it compare the speed at which different algorithms stabilise as the season progresses. Several studies on the outcomes of football matches have revealed a clear research gap: no study has systematically investigated the evolution of Accuracy, AUC-score or probabilistic calibration for ML and DL models in football match prediction. This dissertation will address this exact gap by examining how MLR, RF, GB and MLP models evolve as more data becomes available throughout the season. Additionally, it will investigate the stability of football match prediction performance across first- and second-tier European leagues.

3 Methodology

This chapter outlines the methodological framework used to evaluate the temporal evolution of predictive performance in football match outcome prediction. This section will explain how suitable the data sources, preprocessing procedures, model designs, training pipeline and evaluation strategy are for predicting football match outcomes as a multiclass classification problem.

3.1 Research Design

This dissertation employs an empirical, quantitative research design. Four predictive modeling approaches were implemented: a MLR model as a linear baseline, two non-linear tree-ensemble models (RF and GB) and a MLP as a DL method. To ensure comparability across models, default hyperparameter settings are used.

The models were evaluated within a multi-league setting covering six professional football leagues across two competitive tiers. To assess predictive model behaviour under contrasting structural conditions, the 2019/2020 and 2023/2024 seasons were selected. While the 2023/2024 season represents a structurally stable football season, the COVID-affected 2019/2020 season introduced substantial irregularities. This contrast allows for examining how external disruptions influence volatility and time-to-stable prediction.

Within each season, a time-based split structure is applied to analyse temporal performance dynamics.

The models were trained using match-level data only. Player-tracking information was not incorporated, as high-quality and consistently available tracking datasets were not accessible for the leagues and seasons under investigation.

3.2 Data Description

To conduct a proper investigation of how model prediction performance evolves, it is crucial to obtain high-quality data and apply thorough data cleaning procedures.

3.2.1 Data Sources

The data used in this dissertation consist of match-level information that combine structured match outcomes with corresponding betting odds, team-strength indicators, team-form measures, full-time and half-time scores, as well as dummy variables for the multi-class classification. Table seven in the Appendix provides an overview of all used variables in this modeling framework.

The empirical analysis covers six professional football leagues, categorised into two competitive tiers. The first tier comprises the Bundesliga, La Liga and the Premier League, whereas the second tier includes the 2. Bundesliga, La Liga 2 and the Championship.

All datasets were obtained from the English platform football-data.co.uk, which offers an extensive range of football statistics from international leagues, primarily curated for betting and analytical purposes. This source is widely used in academic and applied sports analytics due to its structured format and long-term data availability.

3.2.2 Variables and Features

To predict football match outcomes using ML and DL models, it is essential to define the target variable (Y) and the predictor variables (X). In this dissertation, the Y is a three-class categorical outcome representing the final match result: Home Win (H), Draw (D), or Away Win (A). A Home Win (H) indicates that the home team has won the match, whereas an Away Win (A) means that the away team was victorious. If the match ends without a winner, the outcome is classified as a Draw (D).

$$Y = H, D, A \tag{1}$$

X, also referred to as features, represent the inputs used by the model to forecast Y. The prediction models learn patterns from these features by extracting meaningful information that contributes to distinguishing between the possible match outcomes.

In this dissertation, the feature set consists of match statistics, historical performance indicators and betting odds. In addition to the features already contained in the dataset, further variables were created through feature engineering. Detailed explanations of the feature engineering process are provided in Section 3.3.1 Feature Engineering. A complete overview of all features used in the model is provided in Appendix A.

3.2.3 Data Cleaning

First, the datasets for the 2019/2020 and 2023/2024 seasons from all six leagues were downloaded from the platform football-data.co.uk.

An exploratory data analysis (EDA) was then conducted to assess data quality. Initially, the raw files contained historical data ranging from the 2015/2016 season up to the most recent seasons, which included several missing values, incomplete rows and inconsistencies. The seasons 2015/2016 until 2018/2019 were removed to retain only the two seasons included in this

study. In addition, inconsistent date formats were standardised to ensure uniform processing across all leagues.

For both seasons, the extent of missing or inconsistent data was minimal, as the datasets were already relatively clean. Only a small number of matches were missing halftime scores. A closer inspection revealed that these matches had been interrupted due to various incidents, which explains the absence of halftime information. To address this issue, missing halftime scores were replaced with zero and an additional dummy variable (Forfeit) was created to flag these matches explicitly. This approach preserves the matches in the dataset while ensuring that the model can account for their irregular nature.

Overall, the final datasets exhibited only minor issues with missing values and incomplete rows what substantially simplified the data cleaning procedures and reduced the need for extensive data correction or row removal.

3.3 Data Processing

After completing the EDA and cleaning the datasets, further preprocessing steps were required to prepare the data for model training.

3.3.1 Feature Engineering

In addition to the variables provided in the raw datasets, several engineered features were created to enhance model performance and capture relevant aspects of team form and match dynamics.

First, season and division identifiers were added and matched with the corresponding matchdays to ensure that each observation could be clearly assigned to the correct league and competitive tier. The date column was also standardised into a consistent format to prevent inconsistencies in subsequent processing steps. Following this, additional performance-related features were engineered. These variables were informed by established practices in football outcome modeling and were inspired by the feature design presented in Atta Mills et al. (2024).

A complete overview and description of all engineered features used in the analysis is provided in Appendix B, including rolling performance and team-strength indicators.

3.3.2 Train-Test Split and Time-Based Splits

Since this dissertation investigates how predictive performance changes as more training data become available, it was essential to implement time-based splits within each league's season. Instead of applying random shuffling or a conventional 80/20 train–test split, the data were divided chronologically into four successive splits (T1–T4). This design ensures that models are always trained on past matches and evaluated on future matches, preventing temporal leakage and enabling an assessment of how increasing information affects predictive performance.

Each split contains an increasing proportion of the season, meaning that the amount of training data grows from T1 to T4. Although the number of splits was kept consistent across leagues, their absolute lengths differ due to varying season structures. Consequently, the time splits represent equivalent relative portions of each season, rather than fixed absolute matchday counts. For transparency, the specific split definitions for each league are as follows:

Table 1: Time-Based Split Definitions for Each League

League	**Matchdays**	**T1**	**T2**	**T3**	**T4**
Bundesliga	34	MD 1–7	MD 1–14	MD 1–21	MD 1–28
2. Bundesliga	34	MD 1–7	MD 1–14	MD 1–21	MD 1–28
La Liga	38	MD 1–7	MD 1–14	MD 1–21	MD 1–28
Premier League	38	MD 1–7	MD 1–14	MD 1–21	MD 1–28
La Liga 2	42	MD 1–8	MD 1–16	MD 1–24	MD 1–32
Championship	45	MD 1–9	MD 1–19	MD 1–28	MD 1–36

Although La Liga 2 and the Championship include slightly more training observations per split, the differences are not large enough to introduce systematic bias. What matters for comparability is that each split reflects a comparable season phase across leagues (early, early-mid, mid, late).

3.3.3 Handling Class Imbalances

An important consideration in multiclass classification tasks is the handling of class imbalance, as ignoring this issue can lead to biased outcome probabilities and misleading predictive performance. Class imbalance is a common problem in football match outcome prediction because the distribution of Home Wins (H), Draws (D) and Away Wins (A) is inherently unequal. In

particular, Home Wins (H) are overrepresented in the data, while Away Wins (A) and especially Draws (D) occur less frequently (Atta Mills et al., 2024).

This imbalance is problematic because predictive models learn directly from outcome frequencies and therefore implicitly assume that frequent outcomes are more important. As a result, models are rewarded for predicting the majority class H and are less penalised for misclassifying rare outcomes such as D and A. Consequently, model predictions become biased towards the majority class. This bias leads to misleading Accuracy values, as a model may achieve high Accuracy by overpredicting H while largely ignoring minority outcomes, resulting in poor true predictive quality.

In addition, the underrepresentation of D and A causes these outcomes to receive systematically lower predicted probabilities, leading to poor probability calibration and overconfident predictions. These issues are particularly critical for this dissertation, which focuses on probabilistic forecasts and prediction stability over time. Unaddressed class imbalance can artificially inflate early-season volatility, distort stability thresholds and ultimately render probability estimates unreliable for decision-making.

To address class imbalance, each model incorporates a RandomOverSample within its preprocessing pipeline to oversample the minority classes. RandomOverSampler was chosen because it duplicates existing minority samples until all classes reach the same size, meaning that no synthetic data is generated. This approach offers a simple and computationally efficient solution for handling imbalanced datasets and is well suited for structured, low-dimensional data such as the match-level information used in this dissertation. Additionally, the RandomOverSampler is well suited for the data used in this dissertation because the dataset is relatively small and the features are aggregated and partially categorical.

Alternative imbalance-handling approaches such as SMOTE or class weighting were considered but not applied in order to maintain a consistent preprocessing strategy across all evaluated models. Using different imbalance-handling strategies for different algorithms would have made the comparison of predictive performance less clear, as performance differences could then arise either from the model itself or from the imbalance-handling technique. To avoid this confounding effect and ensure comparability, the RandomOverSampler was applied uniformly across all models.

3.3.4 Scaling Encoding

Furthermore, it was important to preprocess numerical and categorical features separately to avoid data leakage, ensure correct handling of missing values and improve overall model performance.

For this purpose, a ColumnTransformer was used to apply the appropriate transformations to each feature type. Numerical features were median-imputed using a SimpleImputer and subsequently scaled with a StandardScaler. Categorical features were imputed with the most frequent value using a SimpleImputer and then one-hot encoded using a OneHotEncoder.

The only difference in preprocessing between the four models is that the GB and MLP models use one-hot encoding without a frequency threshold. This is because these models can efficiently handle higher-dimensional sparse feature representations, making it unnecessary to group infrequent categories.

Applying these preprocessing steps within a unified pipeline also ensures that all transformations are learned exclusively from the training data, thereby preventing information leakage into the test set.

3.4 Model Specification

In this subsection, each of the selected models is introduced by outlining its core working principles, followed by a justification for its inclusion in the modeling framework applied in this dissertation.

3.4.1 Multinomial Logistic Regression (MLR)

The baseline model for dealing with the multi-class classification is the MLR.

This model is a linear classifier which models the log-odds of each outcome class and models the relationship between input features and their probability of belonging to a particular class (Atta Mills et al., 2024; Choi et al., 2023).

The model is trained by maximising the likelihood of the correct class labels, adjusting its weights so that the predicted probability for the true class becomes as high as possible. During prediction, the class associated with the highest softmax probability is selected as the model's output.

The MLR model serves in this study as an interpretable baseline against which more complex models can be compared.

3.4.2 Random Forest Classifier (RF)

As a non-linear alternative to MLR, this dissertation employs a RF classifier, a widely used tree-based ensemble method. A RF builds a large number of decision trees and aggregates their predictions to obtain a more robust and accurate classifier (Choi et al., 2023).

Each tree is trained on a bootstrap sample of the training set and at each split a random subset of features is considered. This combination of bootstrapping and feature randomness reduces correlation between trees and helps prevent overfitting (Atta Mills et al., 2024; Baboota & Kaur, 2018).

Given its ability to model non-linear feature interactions, handle high-dimensional input spaces and resist overfitting, the RF serves as a strong non-linear benchmark within the modeling framework.

3.4.3 Gradient Boosting (GB)

GB is another ensemble decision tree technique, but unlike the RF Classifier, which relies on bagging, GB is based on sequential boosting.

In this approach, trees are trained one after another with each new tree learning to correct the errors made by the previous models (Baboota & Kaur, 2018; Bunker et al., 2023). Rather than using random subsets of the data, GB trains each classifier on the entire dataset, updating the model iteratively so that the ensemble progressively improves (Atta Mills et al., 2024; Choi et al., 2023).

By adding weak learners in a stage-wise fashion and optimizing a differentiable loss function, GB reduces bias and enhances predictive accuracy, making it a powerful model for multiclass classification tasks.

3.4.4 Multi-Layer Perceptron (MLP)

The MLP is used as the fourth model in this dissertation and represents the deep learning component of the modeling framework.

The MLP model used in this study consists of two hidden layers of sizes 64 and 32 respectively and is using rectified linear unit (ReLU) activations and the Adam optimiser. In addition, a conservative learning rate of 0.001 is applied combined with a L2 regularisation. The model is using a mini batch size of 64 and the network is trained for up to 500 iterations. This kind of configuration is a stable and moderate complex architecture which is suitable for structured, tabular match-level data. These model configuration represents a generic rather than aggressively optimised hypertuning.

The reason why the MLP model was included is to evaluate whether DL models can capture non-linear feature interaction more effectively than classical ML models in structured, tabular match-level data.

3.5 Hyperparameter Tuning

In this research, all models were trained using their default hyperparameter configurations. This decision ensures a neutral and fair comparison between modeling approaches, as individually fine-tuning each model would introduce comparison bias and obscure the temporal effects that this study aims to analyse. By relying on default settings, the evaluation focuses on the intrinsic behaviour of the models rather than on performance gains that might arise from model-specific optimisation.

While this approach preserves comparability, it also means that the reported results may not reflect the maximum achievable performance of each model. In general, hyperparameter tuning is used to enhance predictive accuracy and calibration by adjusting learning-related parameters such as tree depth, learning rate or regularisation strength. However, applying different optimisation strategies to each model would compromise the methodological consistency required for this study. To maintain a controlled evaluation framework, all models were therefore trained under identical conditions, using the same time-based train–test splits and the same preprocessing and evaluation pipelines.

3.6 Evaluation Strategy

After outlining the mechanisms and specifications of the models used in this study, it is essential to describe the evaluation strategy applied to assess their predictive performance and the metrics used for this purpose.

3.6.1 Performance Metrics

In order to assess the predictive performance of the different models, it is essential to define the criteria that characterize a good football outcome prediction and to specify how these criteria are captured by the chosen evaluation metrics.

On the one hand, a good prediction model must achieve a high level of classification Accuracy. The Accuracy metric measures the overall correctness of predictions by quantifying the proportion of matches for which the model correctly identifies the categorical outcome (Home Win, Draw, Away Win) (Atta Mills et al., 2024; Choi et al., 2023; Bunker et al., 2023).

Accuracy is calculated as the number of correctly predicted outcomes divided by the total number of predictions (Atta Mills et al., 2024; Choi et al., 2023; Wheatcroft & Sienkiewicz, 2021):

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i) \tag{2}$$

To address the limitations of Accuracy-based evaluation, this study also uses the AUC. Although AUC is originally defined for binary classification, it is widely applied to multi-class settings because of its strong ability to assess a model's discriminative power across classes (Atta Mills et al., 2024). In this study, AUC is computed using a one-vs-rest extension, where a separate ROC curve is created for each outcome class and the results are macro-averaged.

The Receiver Operating Characteristic curve (ROC) itself visualises the trade-off between the true positive rate and the false positive rate across different classification thresholds (Choi et al., 2023). A high AUC score indicates that the model is effective at ranking and distinguishing between the outcome classes, even in the presence of class imbalance.

As this dissertation focuses on probabilistic predictions, LogLoss is used to evaluate the quality of predicted probability distributions. Unlike Accuracy, LogLoss incorporates the full predicted probability vector and penalises overconfident but incorrect predictions more strongly, making it well suited for multi-class football outcome forecasting.

LogLoss is equivalent to the cross-entropy between the true outcome distribution and the predicted probabilities. Lower LogLoss values indicate better calibrated and more reliable probability estimates.

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \tag{3}$$

where N denotes the number of observations, M the number of outcome classes, y_{ij} the true class indicator and p_{ij} the predicted probability.

3.6.2 League-Specific Evaluation

The research question is evaluated by model performance comparison across different leagues and competitive tiers.

This enables the assessment of how quickly the models converge toward stable predictive performance within distinct competitive environments. Such comparisons help identify whether league-specific characteristics affect the speed and reliability with which models learn accurate patterns over time. The same rationale applies to the comparison between first- and second-tier competitions.

The convergence of model performance is assessed using the four time-based splits defined in Section 3.3.2, allowing performance stability to be examined across successive phases of the season.

3.6.3 Season-Specific Evaluation

To account for temporal variation, model performance is evaluated separately for the 2019/2020 and 2023/2024 seasons. This allows an assessment of whether predictive performance generalises across different seasonal contexts.

The 2019/2020 season is included due to the disruptions caused by the COVID-19 pandemic, which affected match schedules, league structures and home-advantage dynamics. Comparing this season with the structurally regular 2023/2024 season enables an evaluation of how exceptional conditions influence predictive performance and model stability.

4 Results

In this chapter, the empirical results of the predictive modeling framework are presented and analysed in detail. The section begins by outlining the overall performance of the four models

across all leagues and seasons, followed by league-specific insights that highlight differences in predictability across competitions. Subsequently, the results are contrasted at the tier level to assess systematic distinctions between first- and second-tier leagues. A cross-season comparison evaluates the stability of model performance between the Covid-affected 2019/2020 season and the more recent 2023/2024 season.

The chapter further examines how increasing training data size influences predictive accuracy and model stability over the course of the season. Finally, the models are compared directly to assess their relative strengths and weaknesses. Together, these analyses provide a comprehensive overview of predictive performance dynamics across leagues, tiers, seasons and modeling approaches.

4.1 Overall and Comparative Model Performance

This overall model performance comparison is essential for establishing a robust baseline understanding of how the evaluated models perform across leagues, tiers and seasons.

In terms of Accuracy, the best-performing models in each season are clearly dominated by tree-ensemble approaches. As shown in Tables 2b, these models lead across all six leagues in the 2023/2024 season. Specifically, the RF model achieves the highest Accuracy in the 2.Bundesliga and La Liga 2, while GB emerges as the top performer in the remaining four leagues (see Table 2a).

A similar pattern emerges in the 2019/2020 season. Tree-ensemble models dominate once again, with RF achieving the highest Accuracy in four out of six leagues and GB in the Premier League. The only deviation from this pattern are observed in La Liga, where the MLR model performs best.

The consistent advantage of tree-ensemble models derives from their ability to capture non-linear relationships, interaction effects and heterogeneous patterns within football data. These capabilities enable such models to represent contextual facts, such as form dynamics, team-strength asymmetries and situational variables far more effectively than linear or neural baseline models.

Table 2: Best-Performing Models by Accuracy Across Seasons

(a) 2019/2020 Season

League	Model	Accuracy
2. Bundesliga	RF	0.7685
Bundesliga	RF	0.7188
Championship	RF	0.7680
La Liga	MLR	0.7458
La Liga 2	RF	0.7745
Premier League	GB	0.7528

(b) 2023/2024 Season

League	Model	Accuracy
2. Bundesliga	RF	0.7368
Bundesliga	GB	0.7580
Championship	GB	0.7550
La Liga	GB	0.7805
La Liga 2	RF	0.7488
Premier League	GB	0.7438

Examining model performance on the AUC metric for the 2019/2020 and 2023/2024 seasons reveals notable differences in the hierarchy of best-performing models.

In the 2019/2020 season, model dominance is split between the MLR model and the GB model (see Table 3a). This pattern contrasts with the more uniform results observed in the 2023/2024 season, where GB dominates in five out of six leagues. The only exception is the 2. Bundesliga, where the MLR model again emerges as the top performer (see Table 3b).

Table 3: Best-Performing Models by AUC Across Seasons

(a) 2019/2020 Season

League	Model	AUC
2. Bundesliga	GB	0.9085
Bundesliga	GB	0.9038
Championship	GB	0.9105
La Liga	MLR	0.9043
La Liga 2	MLR	0.9235
Premier League	MLR	0.9020

(b) 2023/2024 Season

League	Model	AUC
2. Bundesliga	MLR	0.8795
Bundesliga	GB	0.9058
Championship	GB	0.8995
La Liga	GB	0.9328
La Liga 2	GB	0.8843
Premier League	GB	0.9023

The Logloss metric is showing the same model tendencies like in Accuracy and AUC. In the 2019/2020 season, the RF model is dominant in five out of six leagues, while the MLR model is performing best in La Liga 2 (see Table 4a). In the 2023/2024 season, this pattern is even

stronger. Here does the RF model dominate in all six leagues as the best model (see Table 4b).

Table 4: Best-Performing Models by Logarithmic Loss Across Seasons

(a) 2019/2020 Season

League	Model	LogLoss
2. Bundesliga	RF	0.6758
Bundesliga	RF	0.6565
Championship	RF	0.6515
La Liga	RF	0.6575
La Liga 2	MLR	0.6365
Premier League	RF	0.6313

(b) 2023/2024 Season

League	Model	LogLoss
2. Bundesliga	RF	0.6875
Bundesliga	RF	0.6670
Championship	RF	0.6523
La Liga	RF	0.5758
La Liga 2	RF	0.6928
Premier League	RF	0.6560

After examining the best-performing models across the key metrics like, Accuracy, AUC and LogLoss, it becomes evident that tree-ensemble methods such as RF and GB are the most stable models across seasons and leagues. While tree-ensemble models consistently dominate, the MLR is the only other model which achieve leading positions in isolated cases. However, the MLP model does not play a pivotal role in this metrics, what could be caused by the restricted amount of data used in this study. Future studies could examine how this model performs if datasize increases immensively. This pattern indicates that the MLP and MLR models are more sensitive to league structures and season-specific dynamics, resulting in greater variability in their performance.

Several dataset characteristics help explain these performance patterns. Leagues with more stable competitive dynamics, such as the Bundesliga, tend to favour ensemble models, whereas more unpredictable leagues allow simpler models to remain competitive. The COVID-affected 2019/2020 season further amplified volatility, which contributes to differences between seasons.

4.2 League-Specific Performance Comparison

Predictive performance varies substantially across leagues. First-tier competitions such as the Bundesliga, La Liga and Premier League generally exhibit smoother and more consistent trajectories across the four time splits (T1–T4), whereas second-tier leagues like 2. Bundesliga, the English Championship and La Liga 2 display markedly stronger early-season volatility and

larger fluctuations before stabilising later in the season (see C.1 & C.2 in Appendix).

4.3 Tier- and Season Level Comparison

In this dissertation, predictive stability is defined as the point in the season at which a model's performance metric exhibits only marginal variation across subsequent time splits, indicating that additional training data yields diminishing improvements. A prediction is considered stable once the absolute change in the evaluated performance metric between consecutive time splits falls below a predefined volatility threshold. Hereby, volatility refers to the variability of the performance metrics (Accuracy, AUC and LogLoss) of the models across the four consecutive time splits (T1-T4) within a fixed league–season–model combination. To enable a compact comparison of predictive stability across tiers and seasons, a time-to-stability measure is defined based on a volatility threshold. Specifically, a stability threshold is set at the 25th percentile of the empirically observed volatility values across all league–season–model combinations. The 25th percentile was chosen as a conservative threshold that identifies stabilisation only once volatility falls well below typical early-season fluctuations.

For each tier–season–model combination, the earliest time split (T1–T4) at which predictive performance remains within this stability threshold is recorded as the point of stabilisation. These discrete stabilisation points are then aggregated across observations. The resulting mean time-to-stable split therefore represents the average point in the season at which stable predictive performance is first achieved, with lower values indicating earlier stabilisation.

Table 5 summarises the stability splits for all performance metrics across first- and second-tier leagues. Across every metric, Tier 1 leagues reach mean stable predictive performance earlier than Tier 2 leagues. For example, the mean stability split for Accuracy is 3.21 in Tier 1 compared to 3.46 in Tier 2, while AUC shows a similar pattern (3.25 vs. 3.38). Loss-based metrics likewise indicate lower stability thresholds in first-tier competitions, with LogLoss stabilising at 3.29 in Tier 1 and 3.38 in Tier 2.

A median stability split of four indicates that, in the majority of cases, predictive performance only stabilises in the late-season phase, highlighting that early- and mid-season predictions remain volatile for many tier–model combinations. These differences are further reflected in the percentage of cases stabilising by T2, which is consistently higher in Tier 1 across all metrics. Conversely, Tier 2 exhibits a higher percentage of cases requiring the final split (T4) to

reach stability. Collectively, this evidence demonstrates that second-tier leagues exhibit greater early-season uncertainty and require more matches before reliable predictive performance is achieved.

The observed differences in time-to-stability across tiers are descriptive rather than inferential and are intended to illustrate general patterns rather than establish statistically significant differences (see Appendix D).

Table 5: Time-to-stability statistics by season and performance metric

Tier	Metric	Mean stability split	Median stability split	% stable by T2	% require T4
Tier 1	AUC	3.25	4.00	25.00	54.17
Tier 1	Accuracy	3.21	4.00	33.33	62.50
Tier 1	LogLoss	3.29	4.00	20.83	66.67
Tier 2	AUC	3.38	4.00	12.50	58.33
Tier 2	Accuracy	3.46	4.00	12.50	70.83
Tier 2	LogLoss	3.38	4.00	12.50	58.33

The overall shape of model performance trajectories remains comparable across both seasons. As shown in Table 6, however, the 2019/2020 season exhibits noticeably higher early-season volatility than 2023/2024, particularly for Accuracy- and AUC-based evaluations. Moreover, performance variability across modeling approaches and time splits is substantially higher in 2019/2020, indicating increased instability in the COVID-affected season.

Several factors may explain this behaviour: interrupted match schedules disrupted normal form progression, matches played in empty stadiums likely weakened home-advantage effects and heterogeneous player fitness levels following lockdowns may have biased form-based features. Together, these irregularities created a more unstable competitive environment.

Despite this elevated early-season volatility, performance trajectories follow a similar progression in both seasons. All models show clear improvement from T1 to T4, and performance converges at the final split, indicating that once sufficient match data becomes available, predictive behaviour stabilises across seasons (see C.1 & C.2 in Appendix).

This difference in volatility is also reflected in the time-to-stability results. Model performance in the 2023/2024 season stabilises earlier than in 2019/2020 across all metrics (see Table 6

& Appendix E). For example, the mean time-to-stable value for the Accuracy metric is 3.21 in 2023/2024, whereas the 2019/2020 season only reaches stability at 3.46. A similar pattern appears for AUC (3.12 vs. 3.50) and LogLoss (3.25 vs. 3.42).

These results indicate that the 2023/2024 season requires less seasonal information for models to produce reliable predictions, while the 2019/2020 season exhibits higher structural volatility and therefore demands more match data before predictive performance becomes stable.

Table 6: Time-to-stability statistics by season and performance metric

Season	Metric	Mean stability split	Median stability split	% stable by T2	% require T4
2019/20	AUC	3.50	4.00	12.50	70.83
2019/20	Accuracy	3.46	4.00	20.83	79.17
2019/20	LogLoss	3.42	4.00	16.67	70.83
2023/24	AUC	3.12	3.00	25.00	41.67
2023/24	Accuracy	3.21	4.00	25.00	54.17
2023/24	LogLoss	3.25	4.00	16.67	54.17

4.4 Impact of Training Data Size

This subsection addresses a central insight of the analysis: the extent to which the growing volume of training data available over the course of the season influences model performance.

Overall, predictive performance improves consistently from T1 to T4, with the largest gains occurring between T1 and T2 and diminishing improvements from T2 to T3. Between T3 and T4, model performance typically converges (see C.1 & C.2 in Appendix). The metric curves begin to flatten, volatility is greatly reduced and predictive behaviour becomes stable across all models. By the end of the season, the models have effectively learned the competitive environment and additional match data provides only marginal improvements. This convergence indicates that late-season predictions are based on sufficiently rich and stable information, resulting in the most reliable forecasts of the entire season.

While volatility is moderate in first-tier leagues, second-tier leagues display considerably higher fluctuations and a more chaotic performance pattern (see C.1 & C.2 in Appendix). This is particularly evident in the early-season phase, where the performance increase from T1 to

T2 is noticeably larger than in first-tier leagues. As a consequence, second-tier competitions reach predictive stability at a later stage of the season (see D in Appendix). This delayed convergence indicates that second-tier leagues benefit more from additional training data, as they require more matches before the underlying competitive structure becomes sufficiently stable for reliable predictions.

Although some leagues exhibit temporary declines between individual splits, these fluctuations do not contradict the general trend and are attributable to league-specific volatility and shifts in match characteristics. To understand these fluctuations more precisely, an additional volatility analysis was conducted. This volatility is not merely random noise but is systematically driven by shifting class imbalances across the time splits and by league- and season-specific dynamics. This was confirmed through a controlled volatility test in which the test sample size was held constant, demonstrating that instability persists even when sampling effects are removed. In particular, second-tier leagues such as the 2. Bundesliga and La Liga 2 exhibit stronger fluctuations due to pronounced variations in the proportions of outcomes A and D, which contribute directly to higher metric volatility in the early part of the season.

In aggregate, the results clearly demonstrate that increasing training data enhances model performance over time, as the sources of volatility diminish and predictions become progressively more stable.

4.5 Key Findings Summary

In the final subsection of the Results chapter, the key findings of this dissertation are summarised, explained, and explicitly linked to the research questions introduced in Section 1.3. Each research question is addressed directly, drawing on the empirical evidence presented in the preceding subsections.

The first research question was: 'How do performance metrics of established models evolve from the beginning to the end of the season as more matches are played?'. As shown in Section 4.4, predictive performance improves consistently across the time splits from T1 to T4. The early-season phase (T1–T2) is characterised by substantial volatility, while performance converges in T3–T4 in almost every league and season. This pattern is further supported by the analysis in Section 4.4, which demonstrates that the increasing volume of training data is a key driver of this improvement and stabilisation.

It is important to note that a small number of leagues and seasons show performance metrics that do not improve substantially over time and remain close to their initial T1 values. However, these cases represent exceptions and do not contradict the overall trend of increasing predictive performance with growing data availability. These deviations are attributed to league- and season-specific volatility factors, such as disrupted team-form signals, unstable match patterns and fluctuations in class imbalance.

Overall, the findings clearly indicate that model performance generally strengthens as more matches become available over the course of a season.

The second research question asks: 'Are there systematic differences in time-to-stable prediction across league tiers and across seasons?'. As demonstrated in Section 4.3, first-tier leagues reach predictive stability earlier than second-tier leagues. These differences between seasons and tiers are not significant but existent. This pattern appears consistently across all evaluated performance metrics, with first-tier competitions requiring fewer time splits to achieve stable and reliable model performance. In contrast, second-tier leagues stabilise later, reflecting their higher structural volatility and greater uncertainty in early-season match outcomes.

In Section 4.4, similar differences are observed across seasons. Across all performance metrics, the 2023/2024 season reaches its time-to-stable prediction point earlier than the 2019/2020 season. This indicates that the COVID-affected 2019/2020 season required more match data before model performance stabilised. In addition to the class imbalance effects discussed in Section 3.3.3, several contextual factors likely contributed to this delay: interrupted match schedules disrupted normal form progression; matches played in empty stadiums weakened traditional home-advantage structures and heterogeneous player fitness levels after lockdowns may have distorted form-based features.

The last research question examines whether MLR, RF, GB and MLP models adapt differently to seasonal information growth. The results of the overall model performance presented in Section 4.1 show that tree-ensemble models, particularly RF and GB, are the dominant and most stable performers across all leagues and both seasons.

In contrast, the MLP model does not appear as the best performing model in any of these performance metrics. These findings indicate that the four evaluated models adapt differently to seasonal information growth, with tree-ensemble approaches showing the most consistent performance improvements over time.

Taken together, these key findings provide a coherent overview of model behaviour across time, league structures and modeling approaches. They also lay the foundation for the interpretative discussion presented in the following chapter.

5 Discussion

Building on the empirical results presented in the previous chapter, this section develops a detailed interpretation of the key findings and situates them within the broader methodological and theoretical context.

5.1 Interpretation of Key Findings

The first key finding concerns the temporal evolution of model performance. This study shows that predictive performance improves progressively over the course of a season as more training data becomes available. With additional observations, model variance is reduced and underlying performance patterns can be learned more reliably. In contrast, the early-season phase is dominated by substantial noise: team form has not yet stabilised and the limited data leads to high uncertainty in the models' estimates.

The second key finding is that first-tier leagues reach predictive stability earlier than second-tier leagues. This can be explained by the generally lower volatility and stronger team identities in top-tier competitions, which create more consistent and predictable patterns and thus a higher signal-to-noise ratio. Under these conditions, models can learn stable relationships more quickly. In contrast, second-tier leagues exhibit greater unpredictability, higher team turnover and more volatile match outcomes, which slow down the learning process and delay the point at which models achieve stable predictive performance.

The final key finding is that the evaluated models exhibit distinctly different learning behaviours. The most stable performers in this study were the tree-ensemble models, RF and GB. Their robustness can be attributed to their ability to capture non-linearities interaction effects and heterogeneous patterns in the data. In contrast, the MLR and MLP models showed much greater variability. MLR is constrained by its linear structure, which limits its capacity to model complex relationships and leads to weaker performance in dynamic or noisy environments. The MLP, while theoretically able to capture complex patterns, requires substantially more training data

to stabilise. As a result, its performance is more sensitive to league-specific conditions and seasonal irregularities.

5.2 Why Model Performance Differs Across Leagues

The reasons why model performance differs across leagues are multifaceted. A primary factor is the level of competitive balance within a league. When a model is trained on data from a highly balanced league, match outcomes are inherently less predictable, which makes it more difficult for the model to identify stable patterns. As a result, predictive Accuracy stabilises later. In contrast, leagues with clearer disparities in team strength provide stronger and more consistent learning signals, enabling models to achieve reliable performance more quickly.

Furthermore, model performance differs across leagues due to varying levels of volatility. Some leagues are characterised by inherently higher degrees of instability, which makes match outcomes more chaotic and therefore more difficult to predict. This is particularly evident in second-tier leagues, where frequent fluctuations in team strength, squad composition and tactical consistency contribute to a more unpredictable competitive environment.

Another reason why second-tier leagues appear more unstable is the difference in data richness across competitions. First-tier leagues typically offer higher-quality statistics and more consistent feature sets, providing models with clearer and more reliable input signals. This greater data richness contributes to more predictable outcomes and lower volatility. In contrast, second-tier leagues often lack the same depth and accuracy of data, which reinforces their inherent unpredictability and makes stable model learning more difficult.

The final reason why model performance differs across leagues relates to structural and contextual factors. First, leagues vary in the number of matches played, the length of the season and the frequency of break weeks. These differences influence how quickly models can accumulate meaningful information and reach predictive stability. Second, external disruptions, such as those observed in the 2019/2020 season during the COVID-19 pandemic, can significantly distort learning dynamics. Irregular match schedules, altered fitness routines and the absence of spectators introduced unprecedented volatility, which affected the consistency and predictability of match outcomes.

5.3 Practical Implication

The findings of this dissertation offer several practical implications for stakeholders in football analytics, betting markets and predictive modeling.

First, stakeholders in betting markets may benefit from the finding that tree-ensemble models, particularly RF and GB, consistently demonstrate superior stability and performance across leagues and seasons. This makes them the most suitable modeling approaches for the development of operational football prediction systems. Their ability in providing reliable predictions even in volatile environments, like in early-season phases or second-tier leagues, makes them a proper fit for football predictions where robust predictions are essential.

Second, the finding that prediction stability is reached earlier in first-tier leagues than in second-tier leagues is highly relevant for football analysts and practitioners. When developing predictive models for second-tier competitions, this insight underscores the need to incorporate uncertainty estimation mechanisms. Such mechanisms are essential for appropriately handling the substantially higher volatility observed in these leagues, particularly during the early-season phase.

Third, the differences between seasons, particularly the delayed stabilisation in the COVID-affected 2019/2020 season, illustrate how external disruptions can influence predictive reliability. This highlights the importance of continuously monitoring model performance and recalibrating probability outputs during structurally abnormal periods, such as seasons with schedule disruptions, changes in home-advantage dynamics or unusual team-form patterns.

All in all early-season model predictions (T1) are characterised by high volatility and substantial uncertainty. Therefore, both bettors and sports analysts should avoid high-stake decisions or strong reliance on model outputs during this phase and instead interpret early-season predictions conservatively. The results indicate that the most favourable period for applied use is the mid-season phase (T2–T3), where predictive reliability increases rapidly while market efficiency may still be limited, making this phase particularly attractive for value-oriented betting and data-driven analysis. In contrast, late-season predictions (T4) exhibit the highest stability and accuracy but offer limited exploitable value for bettors due to highly efficient markets. Sports analysts, however, can continue to benefit from late-season predictions for evaluation and strategic decision-making due to their high reliability.

5.4 Theoretical Contributions

This dissertation contributes valuable insights into the temporal learning behaviour of ML and DL models within football and sports prediction more broadly. These findings highlight several promising directions for future research.

The first valuable insight relates to the examination of stability thresholds across leagues and seasons. While previous research has primarily focused on season-level evaluations, this dissertation demonstrates that model performance improves systematically as match data accumulates throughout the season and stabilises after league-specific thresholds are reached. These findings add academic value by extending existing theory on temporal learning behaviour, quantifying how performance progression varies across model families and competitive contexts.

Second, the results of this dissertation contribute to cross-league prediction theory by demonstrating that predictive performance varies systematically across tiers and levels of competitive balance. The analyses show that model performance differs between first-tier and second-tier leagues, with first-tier competitions reaching performance stability significantly earlier. By quantifying these differences, the dissertation provides empirical evidence that league structure and competitive balance shape the learning dynamics of predictive models, thereby extending theoretical understanding in this domain.

Third, the findings contribute to model comparison theory by demonstrating clear differences in how the evaluated models perform across leagues, seasons and evaluation metrics. The results show that tree-ensemble models exhibit the most stable and reliable predictive performance, whereas the MLR achieves superior prediction performance under specific metric-, league- and season-related conditions. In contrast, the MLP model does not emerge as a strong performer in this analysis, likely due to its reliance on large training datasets to learn stable representations, combined with the relatively limited data availability in this study. These insights extend theoretical understanding of how different model families adapt to varying data environments in football prediction.

5.5 Limitations

Despite the strengths of this research, several limitations should be acknowledged that may skew or bias the results.

One limitation concerns the reliance on a single data source. Although the study incorporates

six datasets from six different leagues, all originate from the same provider (football-data.co.uk). Using multiple, independent data sources would strengthen the robustness of the analysis and reduce the risk of systematic biases stemming from a specific data-collection methodology.

A related limitation is the restricted feature set available in the chosen dataset. While the research provide with match-history and form features a solid foundation for model training, important predictors such as expected goals, player-level performance indicators and injury information are absent. Incorporating such enriched features could substantially improve model performance and offer a more nuanced representation of team and match dynamics.

Additionally, this research is limited by its geographical scope. Although the study includes leagues from three different countries and two competitive tiers, the analysis remains focused exclusively on European football. As a result, the findings capture predictive performance within European league structures and dynamics, which may not generalise to football environments in other regions such as South America, Africa or Asia. These leagues often differ markedly in style of play, competitive balance, scheduling and data availability, meaning that the conclusions drawn in this dissertation may not fully translate to non-European football contexts.

Another limitation concerns the relatively small variety of models and the absence of extensive hyperparameter tuning. Although this dissertation employs four commonly used approaches, MLR, RF, GB and a MLP model, these represent only a subset of available modeling techniques. More advanced methods, such as modern DL architectures, gradient boosting variants like CatBoost or LightGBM, probabilistic models including Bayesian approaches and hybrid ensemble techniques, were not considered. Incorporating such models could potentially enhance predictive performance and offer deeper insights into how different algorithmic families adapt to football-specific data structures.

Moreover, only default or minimally adjusted hyperparameter settings were used to ensure comparability across models. While this approach supports a fair baseline comparison, it may also inadvertently favour certain models that perform relatively well under default configurations, while disadvantaging those that typically require more intensive tuning to reach optimal performance.

5.6 Recommendations for Future Research

Derived from the limitations identified in this study, several opportunities arise for future research.

First, future work could incorporate more sophisticated modeling approaches and conduct comprehensive hyperparameter tuning to optimise performance. Methods such as advanced DL architectures, modern GB variants (e.g., CatBoost, LightGBM), probabilistic models or hybrid ensemble systems may provide deeper insights into model adaptability and predictive capability.

Second, the limitation of the current feature set could be addressed by integrating richer data sources. Future studies may include player-level information, expected goals metrics, injury and squad-rotation data or transfer-related variables. Incorporating such features would allow researchers to capture a wider range of in-game dynamics, potentially leading to more accurate and context-aware football prediction models.

The final research gap concerns the geographical and temporal limitations of the datasets used in this study. Future research could incorporate additional seasons and expand the analysis to leagues from other continents in order to train models on a broader range of football environments. Such an extension would improve the generalisability of performance results and enable a more comprehensive understanding of how predictive models behave across diverse competitive structures and football cultures.

6 Conclusion

The original purpose of this research was to investigate how predictive model performance in football match outcomes evolves over the course of a season.

Overall, the research demonstrates that predictive performance improves across the season as training data accumulates. In this context, the tree-ensemble models RF and GB consistently outperformed the MLP and MLR models. Furthermore, predictive stability was achieved earlier in first-tier leagues, suggesting underlying structural differences in competitive volatility between tiers.

This research offers several contributions. It advances temporal learning theory by providing quantified stability thresholds across leagues and seasons and it further extends cross-league prediction literature by showing that tier structure and competitive balance meaningfully affect

model learning behaviour. Additionally, this research advances model comparison theory by highlighting differences in prediction stability, calibration and adaptability across algorithmic families.

These insights into model performance evolution can support the selection of appropriate models for predicting match outcomes in the early season. They also indicate that future model developers should incorporate uncertainty estimation when forecasting early season matches or matches in highly volatile leagues. Altogether, the findings of this scientific research provides the foundation for more robust and context-aware predictive systems in football.

Appendices

A Feature Definitions

Table 7: Overview of Predictor Variables and Feature Definitions

Feature Name	Explanation
Matchday	Matchday number within the season.
FTHG	Home team goals scored in the first half.
HTAG	Away team goals scored in the first half.
OddsH	Bookmaker odds for a home win.
OddsD	Bookmaker odds for a draw.
OddsA	Bookmaker odds for an away win.
HomeWinRate_w5	Rolling win rate of the home team over the last 5 matches.
HomeDrawRate_w5	Rolling draw rate of the home team over the last 5 matches.
HomeLossRate_w5	Rolling loss rate of the home team over the last 5 matches.
AwayWinRate_w5	Rolling win rate of the away team over the last 5 matches.
AwayDrawRate_w5	Rolling draw rate of the away team over the last 5 matches.
AwayLossRate_w5	Rolling loss rate of the away team over the last 5 matches.
HomeAttackStrength_w5	Attack strength of the home team based on recent goal scoring.
AwayAttackStrength_w5	Attack strength of the away team based on recent goal scoring.
HomeState_w5	Form state indicator of the home team over the last 5 matches.
AwayState_w5	Form state indicator of the away team over the last 5 matches.
HomeGoalDiffAvg_w5	Average goal difference per match for the home team (last 5 matches).
AwayGoalDiffAvg_w5	Average goal difference per match for the away team (last 5 matches).
HomeWinsMarginGoalRate_w5	Rate of home wins by more than one goal (last 5 matches).
HomeLossesMarginGoalRate_w5	Rate of home losses by more than one goal (last 5 matches).
AwayWinsMarginGoalRate_w5	Rate of away wins by more than one goal (last 5 matches).
AwayLossesMarginGoalRate_w5	Rate of away losses by more than one goal (last 5 matches).
TotalWinRate_w5	Overall win rate over the last 5 matches.
TotalDrawRate_w5	Overall draw rate over the last 5 matches.
TotalLossRate_w5	Overall loss rate over the last 5 matches.
TotalAwayWinRate_w5	Away-only win rate over the last 5 matches.
TotalAwayDrawRate_w5	Away-only draw rate over the last 5 matches.
TotalAwayLossRate_w5	Away-only loss rate over the last 5 matches.
TotalHomeWinRate_w5	Home-only win rate over the last 5 matches.
TotalHomeDrawRate_w5	Home-only draw rate over the last 5 matches.

Continued on next page

Feature Name	Explanation
TotalHomeLossRate_w5	Home-only loss rate over the last 5 matches.
dow	Day of week the match was played (0 = Monday, 6 = Sunday).
match	Match identifier.
HTR_A	Dummy variable: away team leading at halftime.
HTR_D	Dummy variable: match drawn at halftime.
HTR_H	Dummy variable: home team leading at halftime.

B Engineered Features

Table 8: Overview of Engineered Features and Their Definitions

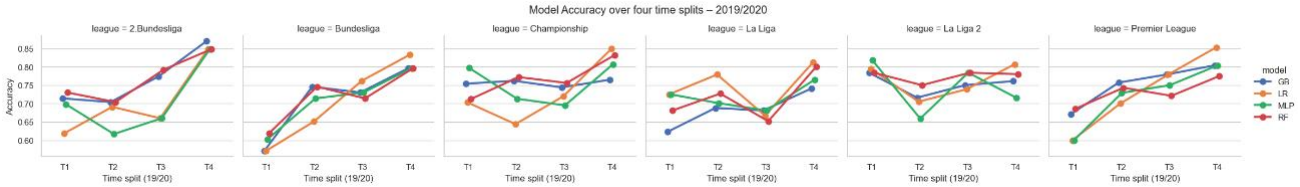
Engineered Feature Name	Explanation
Team-Specific Rolling Performance Rates (last 5 matches)	
HomeWinRate_w5	Rolling win rate of the home team over the last 5 matches.
HomeDrawRate_w5	Rolling draw rate of the home team over the last 5 matches.
HomeLossRate_w5	Rolling loss rate of the home team over the last 5 matches.
AwayWinRate_w5	Rolling win rate of the away team over the last 5 matches.
AwayDrawRate_w5	Rolling draw rate of the away team over the last 5 matches.
AwayLossRate_w5	Rolling loss rate of the away team over the last 5 matches.
Attack Strength Metrics (last 5 matches)	
HomeAttackStrength_w5	Attack strength of the home team based on recent goal scoring.
AwayAttackStrength_w5	Attack strength of the away team based on recent goal scoring.
Form State Indicators (last 5 matches)	
HomeState_w5	Form state indicator of the home team over the last 5 matches.
AwayState_w5	Form state indicator of the away team over the last 5 matches.
Average Goal Differential (last 5 matches)	
HomeGoalDiffAvg_w5	Average goal difference per match for the home team (last 5 matches).
AwayGoalDiffAvg_w5	Average goal difference per match for the away team (last 5 matches).
Margin-of-Victory / Loss Rates (last 5 matches)	
HomeWinsMarginGoalRate_w5	Rate of home wins by more than one goal (last 5 matches).
HomeLossesMarginGoalRate_w5	Rate of home losses by more than one goal (last 5 matches).

Continued on next page

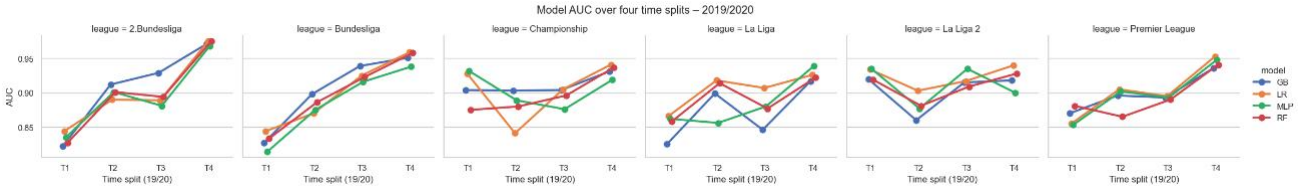
Engineered Feature Name	Explanation
AwayWinsMarginGoalRate_w5	Rate of away wins by more than one goal (last 5 matches).
AwayLossesMarginGoalRate_w5	Rate of away losses by more than one goal (last 5 matches).
Overall Match Outcome Rates (last 5 matches)	
TotalWinRate_w5	Overall win rate over the last 5 matches.
TotalDrawRate_w5	Overall draw rate over the last 5 matches.
TotalLossRate_w5	Overall loss rate over the last 5 matches.
Home- and Away-Specific Outcome Rates (last 5 matches)	
TotalAwayWinRate_w5	Away-only win rate over the last 5 matches.
TotalAwayDrawRate_w5	Away-only draw rate over the last 5 matches.
TotalAwayLossRate_w5	Away-only loss rate over the last 5 matches.
TotalHomeWinRate_w5	Home-only win rate over the last 5 matches.
TotalHomeDrawRate_w5	Home-only draw rate over the last 5 matches.
TotalHomeLossRate_w5	Home-only loss rate over the last 5 matches.
Match-Level and Season-Level Goal Variables	
HomeGoalDifferential	Total season goal difference for the home team.
AwayGoalDifferential	Total season goal difference for the away team.
HomeGoalsFor	Total goals scored by the home team so far in the season.
HomeGoalsAgainst	Total goals conceded by the home team so far in the season.
AwayGoalsFor	Total goals scored by the away team so far in the season.
AwayGoalsAgainst	Total goals conceded by the away team so far in the season.

C League-Level Performance Trajectories

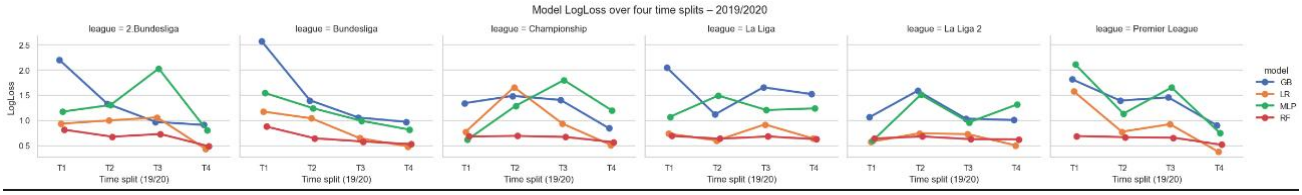
C.1 Accuracy, AUC, and LogLoss Performance Trajectories (2019/2020 Season)



(a) Accuracy



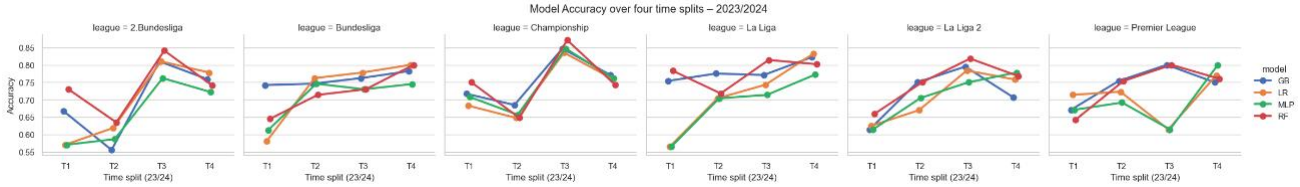
(b) AUC



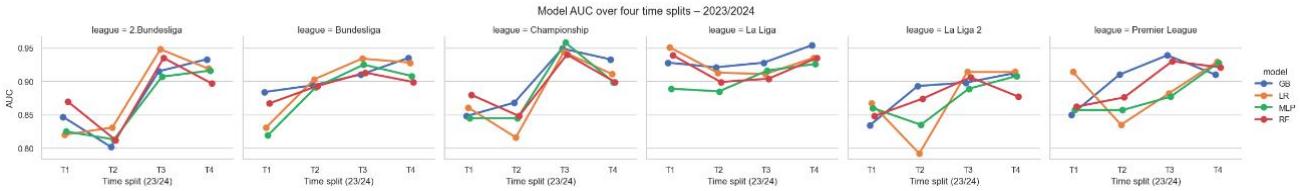
(c) LogLoss

Figure 1: League- and model-level performance trajectories over four time splits (T1–T4) for the 2019/2020 season.

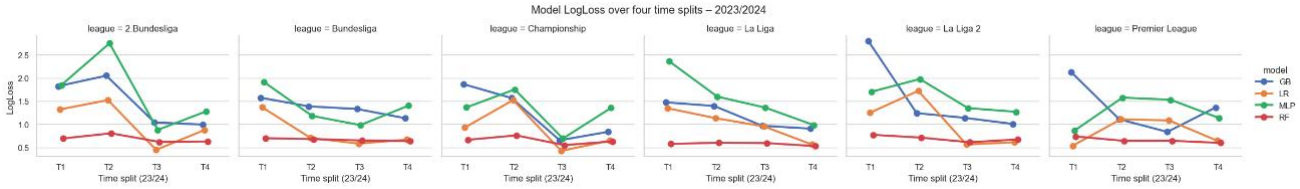
C.2 Accuracy, AUC, and LogLoss Performance Trajectories (2023/2024 Season)



(a) Accuracy



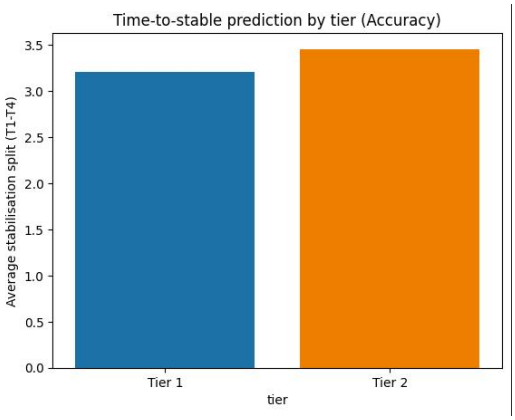
(b) AUC



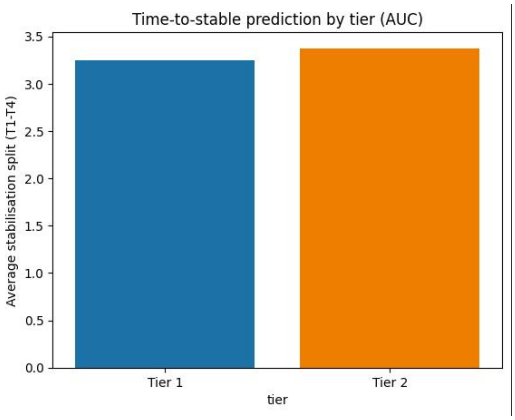
(c) LogLoss

Figure 2: League- and model-level performance trajectories over four time splits (T1–T4) for the 2023/2024 season.

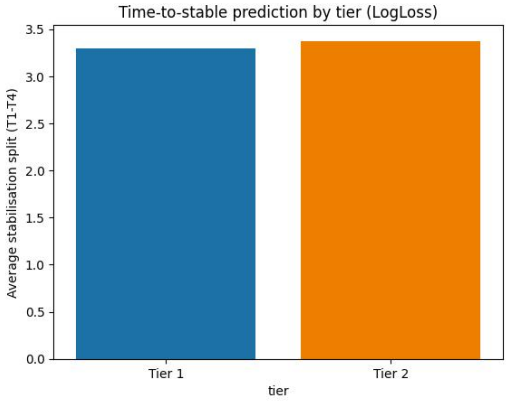
D Tier-level Stabilisation Visualisations



(a) Accuracy



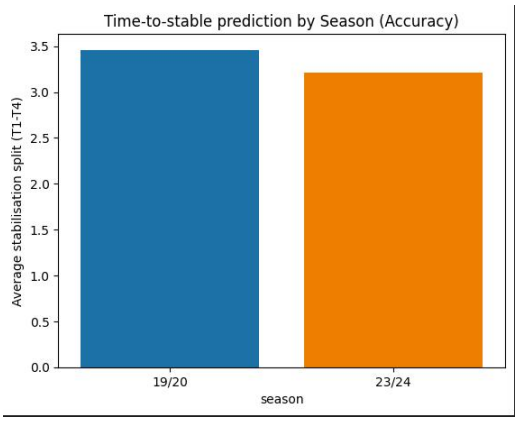
(b) AUC



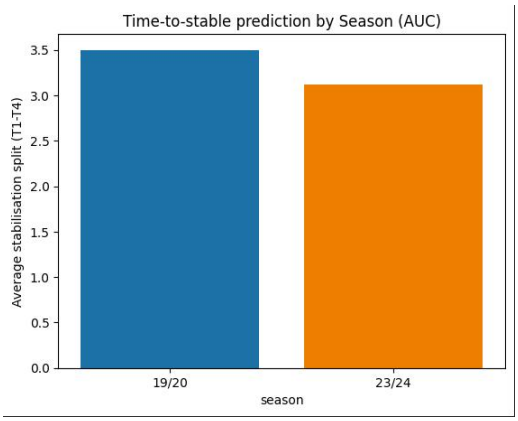
(c) LogLoss

Figure 3: Time-to-stability by competitive tier across performance metrics. Lower values indicate earlier stabilisation (T1 = early season, T4 = late season).

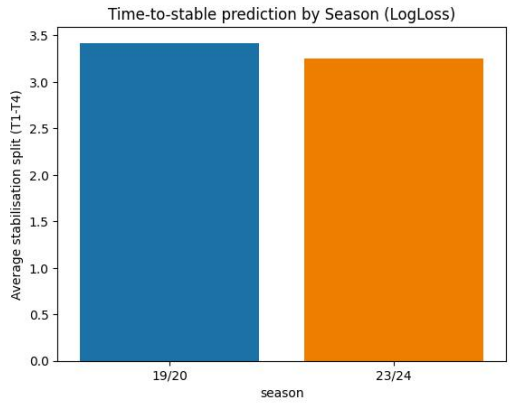
E Season-level Stabilisation Visualisations



(a) Accuracy



(b) AUC



(c) LogLoss

Figure 4: Time-to-stability by season across performance metrics. Lower values indicate earlier stabilisation (T1 = early season, T4 = late season).

Abbreviations

MLR	Multinomial Logistic Regression
RF	Random Forest
GB	Gradient Boosting
MLP	Multi-Layer Perceptron
AUC	Area Under the ROC Curve
ROC	Receiver Operating Characteristic
LogLoss	Logarithmic Loss
ML	Machine Learning
DL	Deep Learning
H	Home Win
D	Draw
A	Away Win
ReLU	Rectified Linear Unit
EDA	Exploratory Data Analysis
Y	Target Variable
X	Predictor Variable

References

Al-Bustami, A., & Ghazal, Z. (2025). From players to champions: *A generalizable machine learning approach for match outcome prediction with insights from the FIFA World Cup.* arXiv preprint arXiv:2505.01902.

Baboota, R., & Kaur, H. (2018). *Predictive analysis and modelling football results using machine learning approach for the English Premier League.* International Journal of Forecasting, 34(4), 741–755.

Baio, G., & Blangiardo, M. (2010). *Bayesian hierarchical model for the prediction of football results.* Journal of Applied Statistics, 37(2), 253–264.

Bunker, R., Yeung, C., & Fujii, K. (2021). *Machine learning for soccer match result prediction.* Sports Analytics Review.

Choi, B. S., Foo, L. K., & Chua, S.-L. (2023). *Predicting football match outcomes with machine learning approaches.* MENDEL Soft Computing Journal.

Constantinou, A. C., & Fenton, N. E. (2012). *Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models.* Journal of Quantitative Analysis in Sports, 8(1).

Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). *pi-football: A Bayesian network model for forecasting association football match outcomes.* Knowledge-Based Systems, 36, 322–339.

Dixon, M. J., & Coles, S. G. (1997). *Modelling association football scores and inefficiencies in the football betting market.* Journal of the Royal Statistical Society: Series C, 46(2), 265–280.

Elo, A. E. (1978). *The rating of chessplayers, past and present.* Arco Publishing.

Groll, A., Hvattum, L. M., Ley, C., Popp, F., Schauberger, G., Van Eetvelde, H., & Zeileis, A. (2021). *Hybrid machine learning forecasts for the UEFA EURO 2020.* European Journal of Operational Research, 294(1), 401–412.

Huang, C., & Zhang, S. (2023). *Explainable artificial intelligence model for identifying market value in professional soccer players*, (arXiv:2311.04599v2)

Hvattum, L. M., & Arntzen, H. (2010). *Using ELO ratings for match result prediction in association football.* International Journal of Forecasting, 26(3), 460–470.

Koopman, S. J., & Lit, R. (2017). *Forecasting football match results in national league competitions using score-driven time series models.* Tinbergen Institute Discussion Paper TI 2017-062/III.

Macrì, R., De Martino, D., Egidi, L., & Torelli, N. (2018). *Bayesian weighted dynamic models for association football prediction.* Journal of the Royal Statistical Society.

Maher, M. J. (1982). *Modelling association football scores.* Statistica Neerlandica, 36(3), 109–118.

Platt, J. C. (2000). *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* Advances in Large Margin Classifiers, 61–74.

Tax, N., & Joustra, Y. (2015). *Predicting the Dutch football competition using public data: A machine learning approach.* IEEE Transactions on Knowledge and Data Engineering.

Wheatcroft, E., & Sienkiewicz, E. (2021). *Calibration and hyperparameter tuning in football forecasting with machine learning.* Proceedings of the International Conference on Machine Learning and Applications (ICMLA).

Wong, A., Li, E., Le, H., Bhangu, G., & Bhatia, S. (2025). *A predictive analytics framework for forecasting soccer match outcomes using machine learning models.* Decision Analytics Journal, 14, 100537.