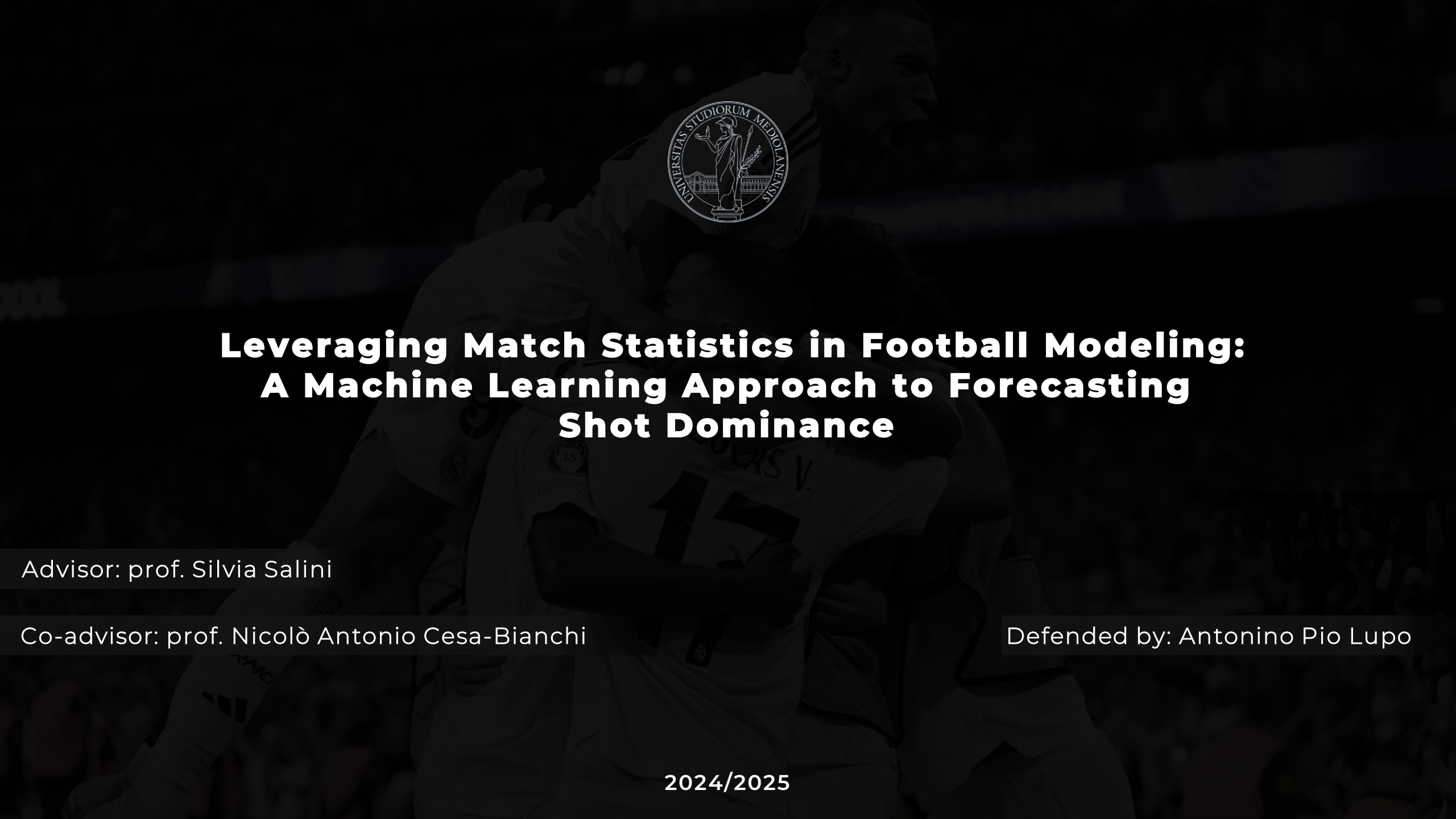


Leveraging Match Statistics in Football Modeling: A Machine Learning Approach to Forecasting Shot Dominance

Advisor: prof. Silvia Salini

Co-advisor: prof. Nicolò Antonio Cesa-Bianchi

Defended by: Antonino Pio Lupo

2024/2025

Agenda

1

RESEARCH CONTEXT

Overview of the **thesis goal**, identified **gaps** in **football prediction literature**, and novel contributions

2

DATASET OVERVIEW

Description of **data acquisition**, **cleaning**, **merging**, **overview**, and **EDA** including target analysis and challenges like biases and imbalances

3

FEATURE ENGINEERING

Methodology for constructing, selecting, and augmenting features from raw data to create **predictive inputs**

4

MODELING FRAMEWORK

Selection of **ML algorithms**, **evaluation metrics**, **calibration techniques**, and integration with **value betting** concepts

5

RESULTS & FINAL REMARKS

Performance comparisons, key **insights**, and **proposed directions** for further research

1

RESEARCH CONTEXT

Research Goal

Employing **pre-match advanced statistics** to design a **machine learning model** that forecasts which team will generate a **higher number of shots** in a football match.

Identified Gaps in Current Football Prediction Research

1. Narrow Focus on Traditional Betting Markets

Most studies still concentrate on predicting **match results** or **goal counts**. While informative, this narrow focus fails to reflect the **growing diversity of betting markets**, where more specific outcomes (e.g., corners, cards, or shots) are increasingly relevant for both bettors and analysts

2. Limited Consideration of Pre-Match Context

A significant portion of the literature models the relationship between **post-match statistics** and outcomes. This approach **reduces predictive reliability** and **limits practical applicability in real-life context**

3. Reliance on Classical Statistical Methods

Many foundational studies in football modeling rely on **traditional statistical techniques**—most notably Poisson models—to estimate scoring rates. While these models are interpretable, they often fail to capture the **complex, nonlinear interactions** present in modern football data

Novel Contributions of This Work

1. Focus on a New Market: Shots

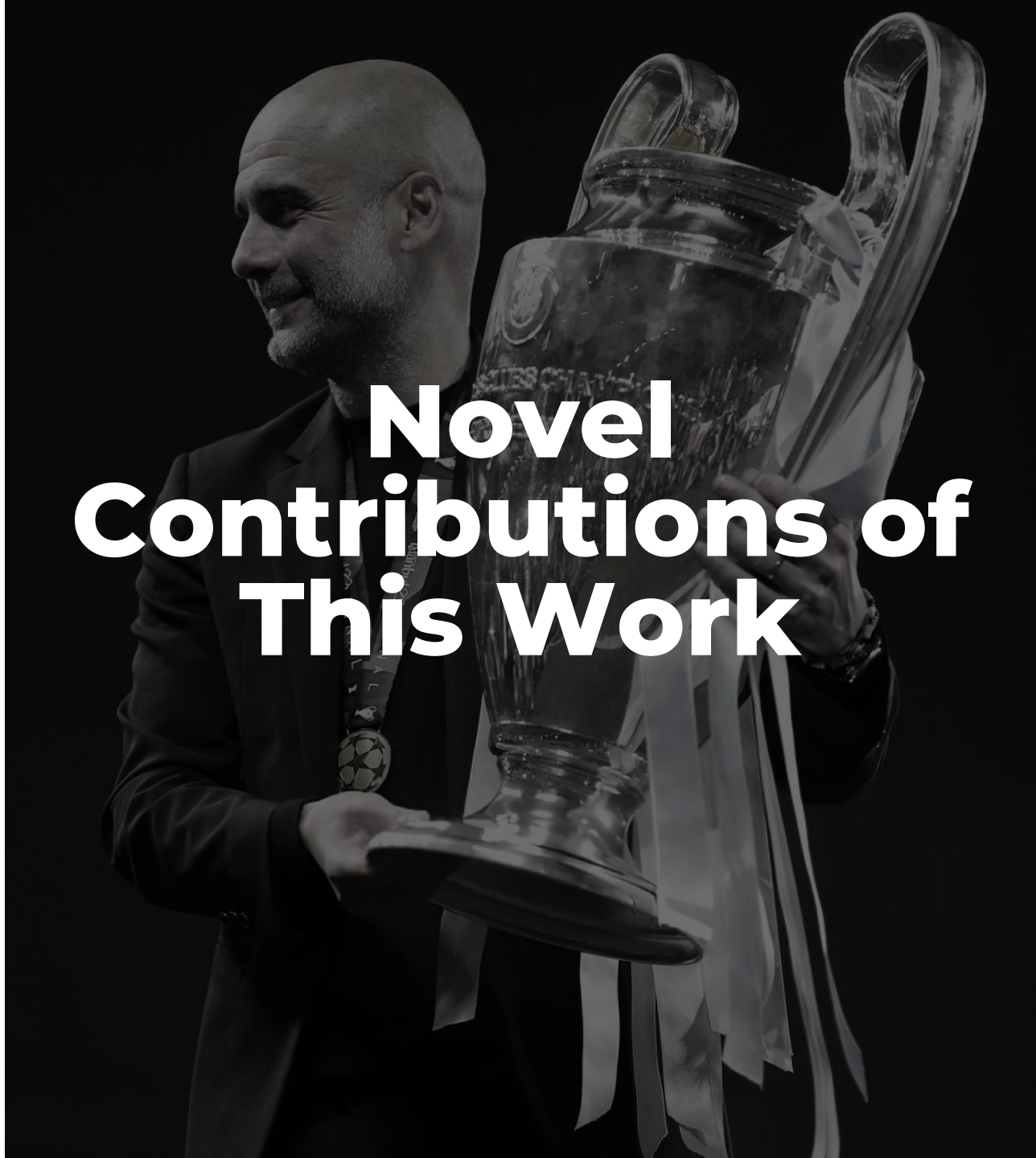
This work explores shot dominance as a predictive target. Unlike traditional outcomes, these markets are **less efficient** and **more linear**, making them **easier to predict** and offering **greater potential value capture** for bettors

2. Pre-Match Contextual Features

Pre-match features are constructed, summarizing team performance in a defined window before each game. This allows the model to generate **predictions for upcoming matches** and **applicable in real-life context**

3. Machine Learning Approach

By using **machine learning**, the model can exploit a large dataset and a rich set of features, combining standard statistics with advanced metrics. This approach captures **nonlinear relationships** that classical models, such as Poisson, may miss



2

DATASET OVERVIEW

Constructing the Dataset

This study integrates match-level data from two public sources: ***FBref*** (for general and advanced metrics) and ***Football-data.co.uk*** (for missing basic statistics and betting odds).

The data was not available as a single unified dataset, therefore it required a **multi-stage process** of **scraping**, **merging** and **cleaning**.

Data Scraping & Acquisition

- Programmatic scraping was performed using the ***worldfootballR*** package in *R* to collect advanced match statistics from ***FBref***
- Data for **shooting**, **passing**, **goalkeeping**, and **goal-creating actions** were retrieved for multiple leagues and seasons
- Complementary data, including **half-time goals**, **corner counts** and **bookmaker odds**, were collected from pre-formatted *CSV* files on **Football-data.co.uk**

Data Cleaning & Merging

- The individual *FBref* tables (shooting, passing, etc.) were first merged into a **single, coherent dataset** using a **synthetic** ***MatchID*** and **composite keys** (*Date, HomeTeam, AwayTeam*)
- Raw data underwent standardization, including **harmonizing column names**, **parsing dates**, and **mapping team names** to a consistent format
- Matches with incomplete statistics were removed to ensure **a fully populated dataset**
- The complementary data from *Football-data.co.uk* underwent the same standardization and harmonization process. The two cleaned datasets were then **merged** to create the **final unified dataset**

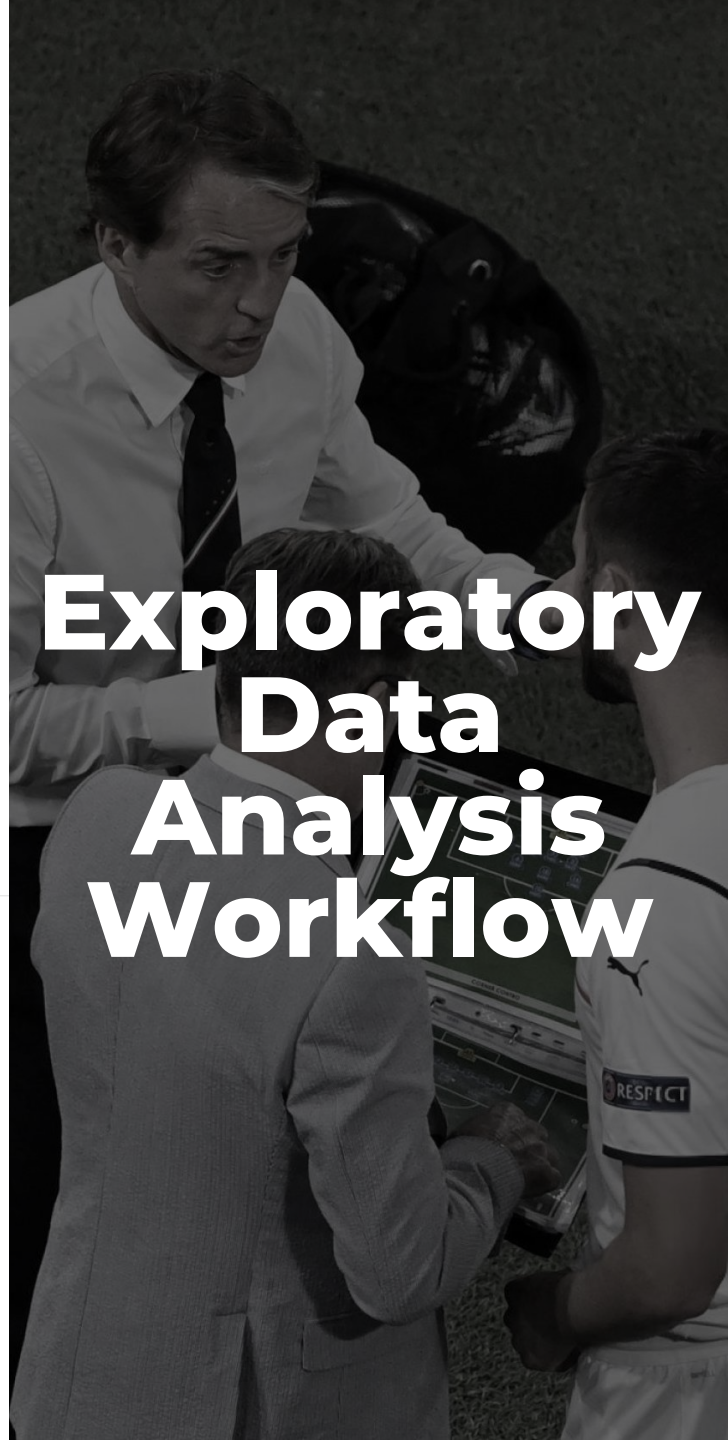
Addition of New Metrics

- New variables were created to enrich the predictive dataset, including:
    - **Match Points**: Points awarded based on match result (*win*=3, *draw*=1, *loss*=0)
    - **Cumulative Standings**: Season-to-date points and league rank for each team
    - **Expected Points** (***xPoints***): Derived via Monte Carlo simulation using both *Expected Goals* (*xG*) and *Post-Shot Expected Goals* (*PSxG*)
    - **Target Variable**: A categorical label (*1*, *X*, *2*) indicating which team took more shots

Dataset Overview

Number of **Matches**	**13,813** unique records
Leagued **Included**	
Seasons **Covered**	from **17/18** to **24/25**
Number of **Statistics**	**63** variables

Exploratory Data Analysis Workflow



Target Variable Analysis

An analysis was conducted on the distribution of **total shots for home** and **away teams**, examining **summary statistics** and **visualizations** to understand data spread, central tendency, and the presence of outliers.

Statistical tests were employed to evaluate the **significancy** of the differences between home and away teams

Correlation Analysis

A **bivariate correlation analysis** was performed to identify which match statistics exhibited the **strongest linear relationships** with the number of shots taken by each team

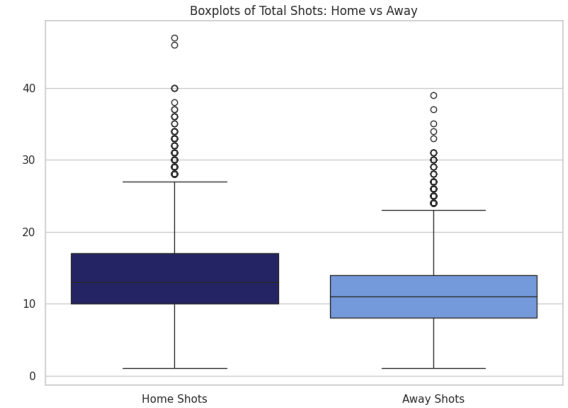
Class Imbalance

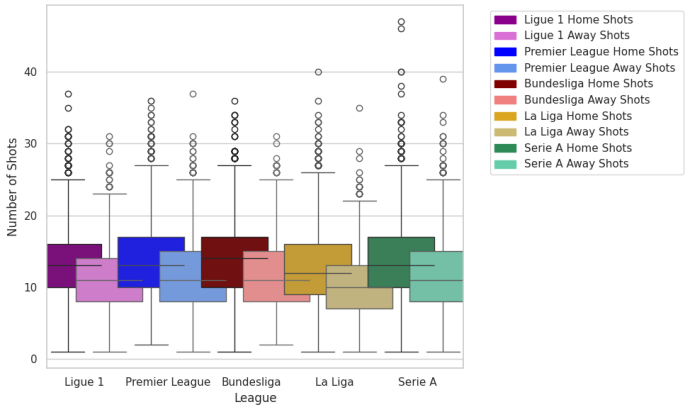
The distribution of the categorical target variable was assessed to quantify **class imbalance**

Dimensionality Reduction & Clustering

Principal Component Analysis was applied to reduce the feature space, followed by ***k-means*** **clustering** on the principal components to identify latent patterns and group matches into distinct profiles.

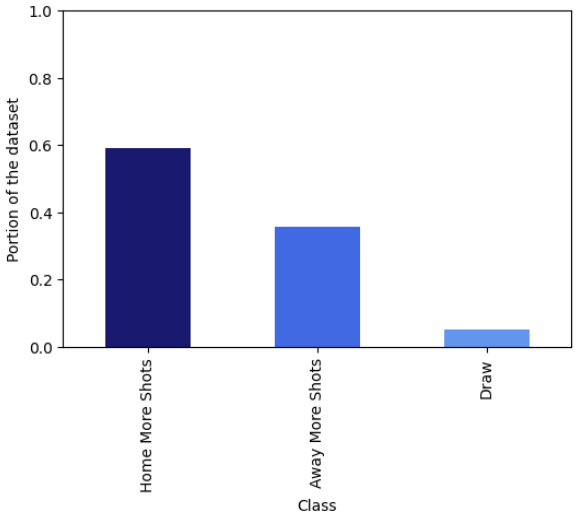
Key Challenges





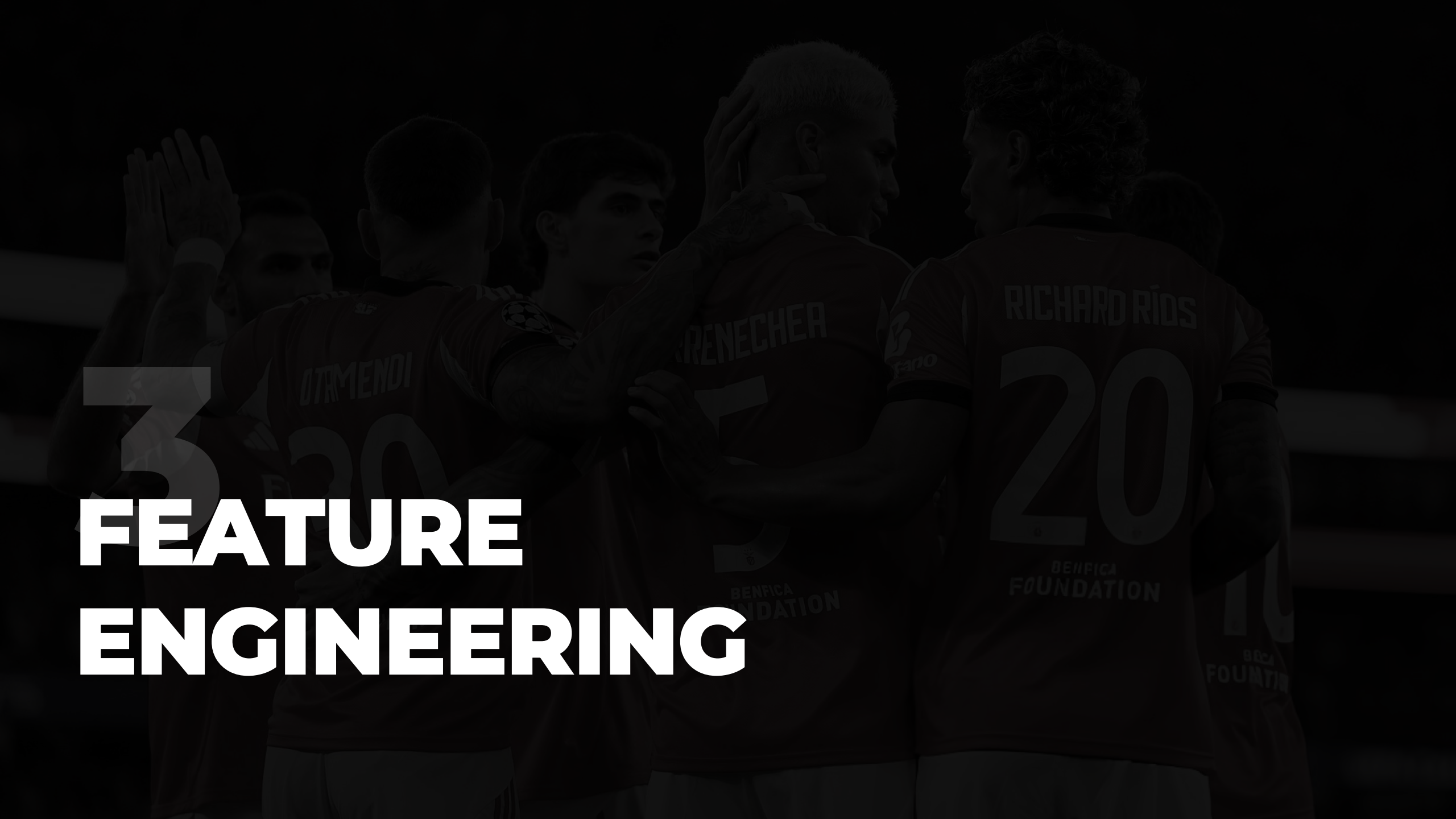
1. Home-Advantage Bias

The dataset exhibits a significant statistical bias where **home teams take on average more shots than away teams**. This may cause the classifier to systematically favor home teams, **compromising its ability to generalize** accurately across all outcome classes



2. Draw Class Imbalance

The outcome where both teams take an equal number of shots is a **very rare event in the data**. This extreme class imbalance presents a challenge for the classifier to learn any meaningful patterns for this specific class



3 FEATURE ENGINEERING

From Raw Data to Model Features

FEATURE CONSTRUCTION

Transforms raw historical match statistics into **pre-match features** using **rolling averages** from preceding matches to inform model predictions before a game begins

FEATURE SELECTION

Employs ***ANOVA*** and ***Random Forest*** importance to identify and retain the most predictive variables, **reducing dimensionality** and **mitigating overfitting risk**

DATA AUGMENTATION

Applies the ***SMOTE*** **technique** to synthetically generate samples for the **minority *away more shots* class**, mitigating model bias caused by class imbalance

Feature Engineering Methodology

1. Feature Construction

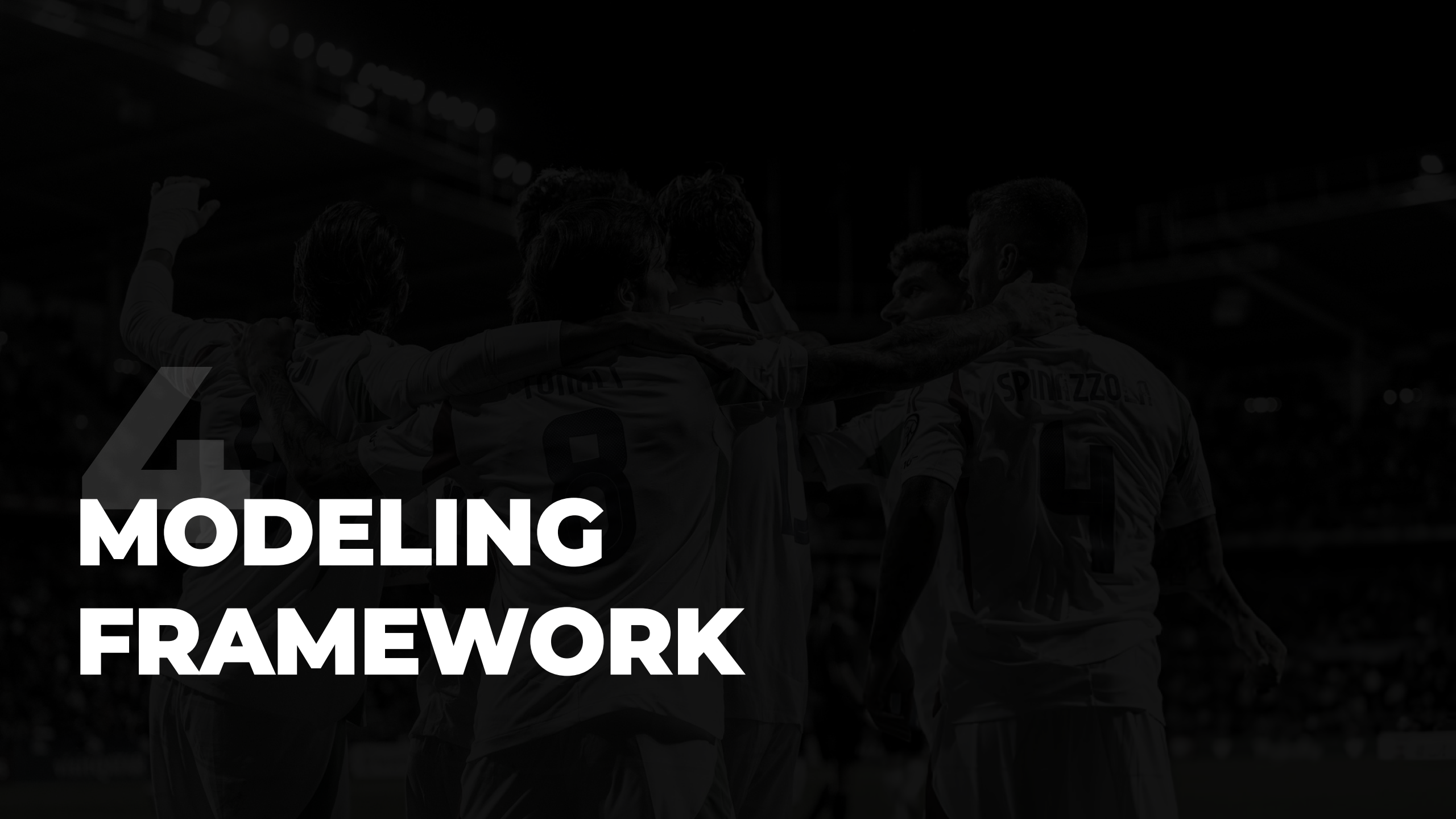
- Features were created as **rolling averages** of every metric over its **last *t* home/away matches**
- **Window size** selection follows ***Berrar et al.*** (***2019***), balancing recent form (*smaller t*) against data robustness (*larger t*)
- Following insights from *Berrar et al.*, a dual-window strategy (***t=6*** and ***t=9***) was implemented to test the best trade-off

2. Feature Selection

- Two distinct techniques were used: ***ANOVA*** (selecting features with $p < 0.05$) and ***Random Forest*** (using an importance threshold)
- The feature set was **progressively reduced** to *75%*, *50%*, and *25%* subsets using these criteria
- **Tree-based models** were exclusively reduced using the **feature importance threshold filter**

3. Data Augmentation

- The ***SMOTE* technique** was applied exclusively to the ***Away More Shots*** class, the largest minority
- The ***Draw*** class was **explicitly excluded** from augmentation due to its **extreme rarity** and **highly stochastic** nature
- This selective approach prevented generating potentially **unrealistic synthetic data**, focusing instead on rebalancing the core home/away dichotomy



4 MODELING FRAMEWORK

Modeling and Evaluation Framework

Model Selection

A diverse set of **machine learning algorithms** was implemented to benchmark performance, ranging from **interpretable linear models** to **complex**, **non-linear ensembles** and **neural networks**

Performance Evaluation

Models were rigorously trained and tested across **different experimental conditions** to identify the **optimal configuration** and compare their predictive reliability on unseen data

Evaluation Metrics

A **comprehensive set of metrics** was employed to **assess model performance**, focusing not only on categorical accuracy but, crucially, on the **quality of the probabilistic forecasts**

Probability Calibration

The reliability of the models' predicted probabilities was assessed, and post-processing techniques were applied to **improve the alignment between forecasted and actual outcome frequencies**

Understand Bookmakers Odds

Bookmakers calculate odds by **converting their probability estimates into decimal form and adding a margin** (*overround*), ensuring total implied probability exceeds 100% to guarantee profit

$$\text{Odds} = \frac{1}{\text{Implied Probability} + \text{Margin}} \qquad \frac{1}{1.50} \approx 0.6667$$

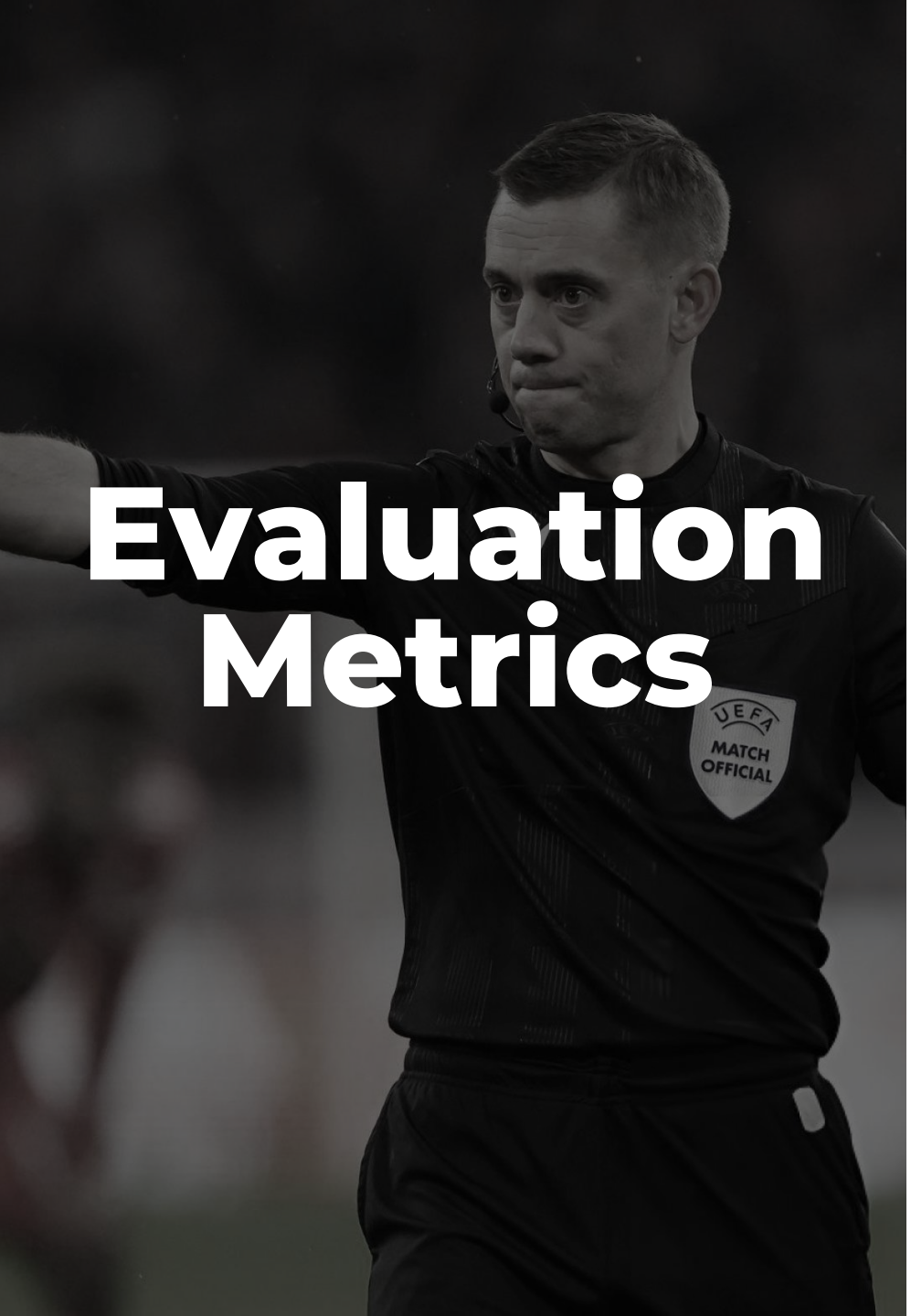
Identify a Value Bet

A **value bet** exists when your **model's predicted probability exceeds** the **probability implied** by the **bookmaker's odds**, indicating positive expected value and long-term profitability

Model's Core Objective

To generate **accurate**, **well-calibrated probability forecasts** that reliably identify when market odds **misprice** true outcome probabilities, enabling sustainable value betting strategies

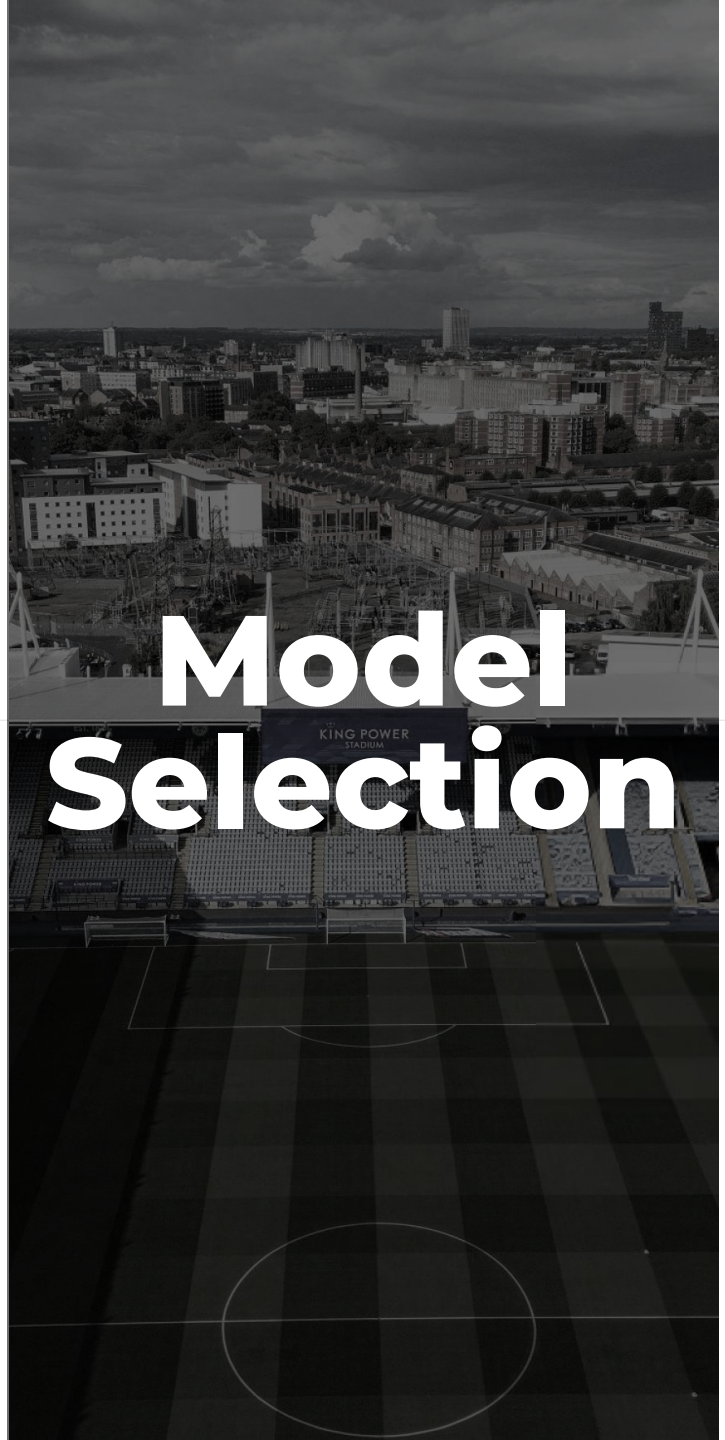
The Anatomy of a Value Bet



Evaluation Metrics

Accuracy	Precision	Recall
$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$

F1 score	Log Loss	Brier Score
$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(\hat{y}_{ij})$	$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\hat{y}_{ij})^2$



Model Selection

LOGISTIC REGRESSION

A **linear model** using the softmax function for multiclass classification, serving as an **interpretable baseline** to benchmark against more complex algorithms

XGBOOST

A **gradient boosting framework** that sequentially builds trees to correct previous errors, optimized for **speed** and **performance** through parallel processing and regularization techniques

RANDOM FOREST

An **ensemble** of **decision trees** that reduces overfitting through bagging and feature randomness, providing **robust performance** on **tabular data** with mixed feature types

NEURAL NETWORKS

A **multi-layer feedforward network** capable of learning **complex non-linear patterns** through backpropagation, offering high capacity for capturing intricate feature interactions in the data

5

RESULTS & FINAL REMARKS

Key Training Insights

Simple Models Performance

The **strong results** from **linear classifiers** indicate that future gains may come from **feature engineering** rather than **increased algorithmic complexity**

Ineffectiveness of Data Augmentation

Applying *SMOTE* **systematically worsened probability calibration**, showing synthetic samples failed to capture real minority class characteristics

Optimal Window Size

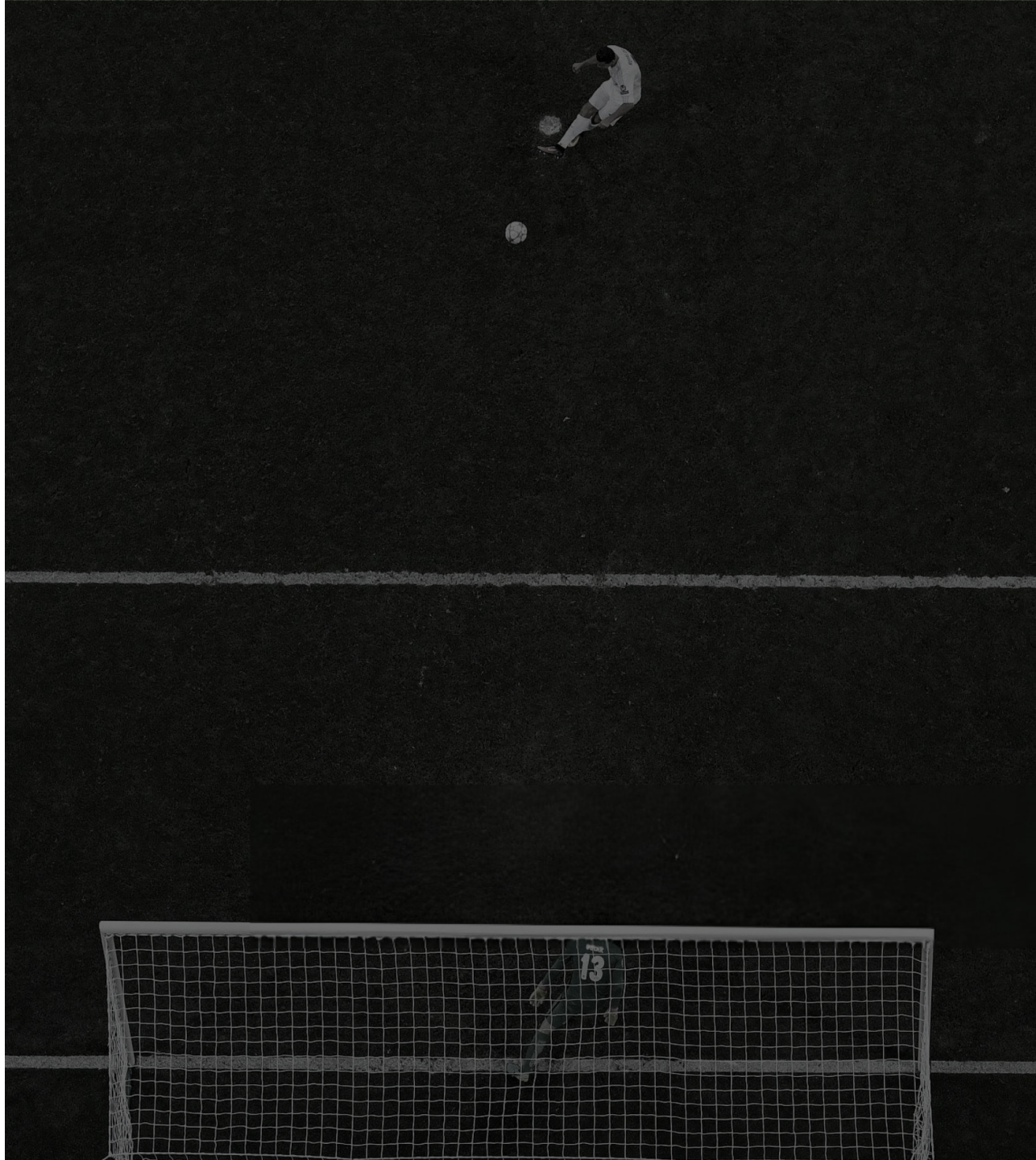
Shorter-term historical data (***t*=6**) consistently **outperformed longer windows**, confirming recent team form is more predictive than extended history

Robustness to Feature Reduction

Feature selection **maintained** or even **improved performance**, revealing significant redundancy and identifying a core set of predictive metrics

Performance Evaluation

Model	Window	Features	SMOTE	Acc	WPrec	WRec	WF1	LogLoss	Brier
Logistic Regression	t=6	FeatImp	No	0.68	0.64	0.68	0.65	0.74772	0.4448
Random Forest	t=6	Full	No	0.67	0.62	0.67	0.64	0.74752	0.4474
XGBoost	t=6	FeatImp	No	0.67	0.63	0.67	0.64	0.74437	0.4453
Neural Network	t=6	25% ANOVA	No	0.68	0.64	0.68	0.65	0.74228	0.4437



Performance Evaluation

Model	Window	Features	SMOTE	Acc	WPrec	WRec	WF1	LogLoss	Brier
Logistic Regression	t=6	FeatImp	No	0.68	0.64	0.68	0.65	0.74772	0.4448
Random Forest	t=6	Full	No	0.67	0.62	0.67	0.64	0.74752	0.4474
XGBoost	t=6	FeatImp	No	0.67	0.63	0.67	0.64	0.74437	0.4453
Neural Network	t=6	25% ANOVA	No	0.68	0.64	0.68	0.65	0.74228	0.4437

Performance Comparison

Strategy	Accuracy	W. Precision	W. Recall	W. F1-Score
Naive Predictor	0.59	0.35	0.59	0.44
Historical Averages	0.65	0.61	0.65	0.61
Betting Odds	0.65	0.63	0.65	0.63
Neural Network	**0.68**	**0.64**	**0.68**	**0.65**

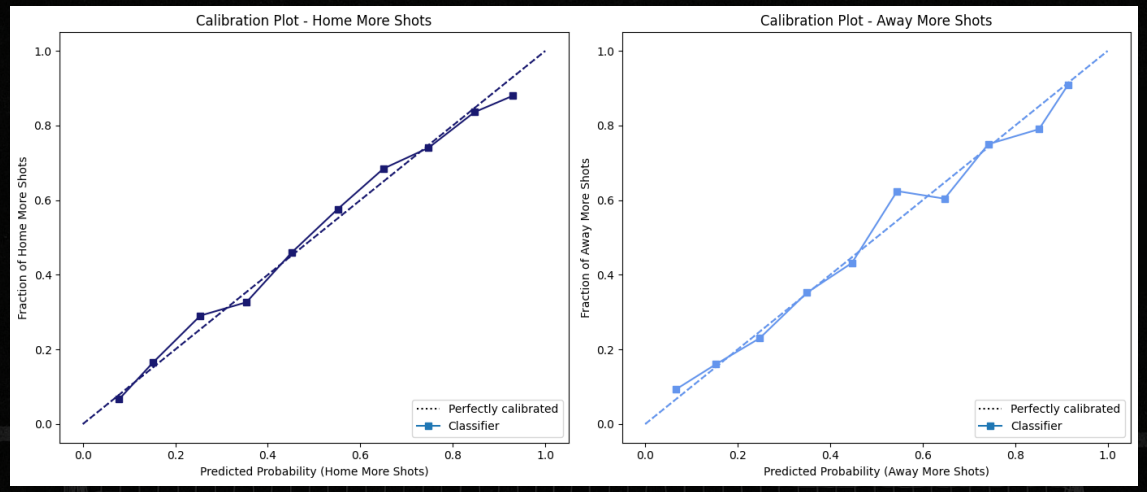
Performance Evaluation

Model	Window	Features	SMOTE	Acc	WPrec	WRec	WF1	LogLoss	Brier
Logistic Regression	t=6	FeatImp	No	0.68	0.64	0.68	0.65	0.74772	0.4448
Random Forest	t=6	Full	No	0.67	0.62	0.67	0.64	0.74752	0.4474
XGBoost	t=6	FeatImp	No	0.67	0.63	0.67	0.64	0.74437	0.4453
Neural Network	t=6	25% ANOVA	No	0.68	0.64	0.68	0.65	0.74228	0.4437

Performance Comparison

Strategy	Accuracy	W. Precision	W. Recall	W. F1-Score
Naive Predictor	0.59	0.35	0.59	0.44
Historical Averages	0.65	0.61	0.65	0.61
Betting Odds	0.65	0.63	0.65	0.63
Neural Network	**0.68**	**0.64**	**0.68**	**0.65**

Calibration Curves



Platt Scaling

Model	Uncalibrated Brier Score	Calibrated Brier Score	Uncalibrated Log Loss	Calibrated Log Loss
Random Forest	**0.4474**	0.6667	**0.7475**	1.0986
XGBoost	**0.4453**	0.4456	**0.7441**	0.7449
Logistic Regression	**0.4448**	0.6667	**0.7477**	1.0986
Neural Network	**0.4437**	0.445	**0.7423**	0.7445

Isotonic Regression

Model	Uncalibrated Brier Score	Calibrated Brier Score	Uncalibrated Log Loss	Calibrated Log Loss
Random Forest	**0.4474**	0.6667	**0.7475**	1.0986
XGBoost	**0.4453**	0.4468	**0.7441**	0.8465
Logistic Regression	**0.4448**	0.6667	**0.7477**	1.0986
Neural Network	**0.4437**	0.4438	**0.7423**	0.75

Calibration Improvement

Parametric Recalibration

- **Platt scaling** was implemented as the parametric calibration method
- This approach applies **logistic regression** to transform **model scores** into **calibrated probabilities**
- The procedure resulted in **degraded performance** across most models, failing to improve reliability

Non-parametric Recalibration

- **Isotonic regression** was employed as the non-parametric calibration technique
- This method fits **a non-decreasing function** to the **predicted probabilities** to better align them with the true outcome frequencies
- The approach caused overfitting and **produced less reliable estimates** than original models

Work Accomplished

Pre-Match Framework

- Developed a **pre-match prediction framework** including advanced metrics using **rolling averages**
- Engineered **historical features** capturing both **team strength** and **recent form** through **dual-window temporal aggregation**
- Implemented comprehensive **feature selection** to reduce dimensionality while maintaining predictive power

ML Implementation

- Trained and compared **multiple classifier families** including Logistic Regression, tree ensembles, and Neural Networks
- Conducted rigorous **hyperparameter optimization** and **probabilistic calibration** for reliable uncertainty estimation
- **Complex models** achieved **marginally better probabilistic calibration**, while **simpler models** maintained **competitive performance** on classification metrics

Performance Insights

- All **developed models** significantly **outperformed baseline strategies**
- Addressed **home advantage bias** through data augmentation
- The extreme rarity of the **draw class** remained a **fundamental challenge**, affecting calibration metrics despite overall strong performance
- Achieved **reliable probability estimates** for *home/away* shot dominance, establishing practical utility for emerging betting markets

Work Ahead

Data Quality

- Expand **historical data coverage** to include more leagues
- Develop methods to **quantify** traditionally **subjective** or **hard-to-measure** factors like fatigue, morale, and lineup
- Inquire into **feature importance** and **multicollinearity** to create more effective and interpretable feature subsets

New Methodologies

- Explore **alternative modeling frameworks** such as regression-based approaches
- Implement **ensemble methods** that combine the strengths of different algorithms for improved stability

Strategy & Validation

- Evaluate **model profitability** through **real-world betting simulations** using historical odds data for shot-based markets
- Develop **practical betting strategies** based on the model's probability estimates for sustainable value betting
- Explore methods to **infer shot dominance odds** from **correlated markets** until direct odds are available


THANK YOU