Contextual Player Evaluation in Football: A Multi-Factor Analysis Approach

by

Prashant Upadhyay

Student number: 2329237

MSC Data Science

School of Mathematics and Computer Science Faculty of Science and Engineering

Module Coordinator: Dr. Andrew Gascoyne

Project Supervisor: Pooja Kaur

Abstract

This research study delves into the realm of football analytics and the application of machine learning algorithms to enhance player performance evaluation and match outcome prediction. The evolution of football analytics is explored, tracing its progression from basic statistics to advanced video analysis and GPS tracking technologies. The report investigates various match analysis systems, including video-based time-motion analysis and GPS systems, highlighting their impact on capturing player movements and tactical patterns. Machine learning algorithms are examined for predicting player ratings, with a focus on the types of algorithms used and the value they bring to stakeholders in football. Additionally, the study delves into the challenges of predicting match outcomes using machine learning models, addressing variables analysed, methodologies employed, and inherent challenges faced. The report concludes by discussing the importance of data quality, model interpretability, and interdisciplinary collaboration in advancing football analytics. Through a comprehensive analysis of player performance metrics and match strategies, this research aims to provide valuable insights for optimizing player performance and enhancing team management in football.

Table of Contents

Abstract 2
Acknowledgements 8
Chapter 1 Literature Review 1
1.1 Introduction 1
1.2 Evolution of Football Analytics 1
1.3 Comparing Football Match Analysis Systems 2
1.3.1 Video-Based Time-Motion Analysis 2
1.3.2 Semi-Automatic Multiple-Camera Systems 3
1.3.3 GPS Systems 4
1.3.4 Impact of System Choice 5
1.4 Utilizing Machine Learning Algorithms for Player Rating Prediction 5
1.4.1 Role of Machine Learning Algorithms 6
1.4.2 Analysed Player Attributes 7
1.4.3 Employed Algorithms 7
1.4.4 Insights and Value 7
1.4.5 Future Directions 8
1.5 Predicting Match Outcomes with Machine Learning 8
1.5.1 Analysed Variables 9
1.5.2 Methodologies Employed 9
1.5.3 Value and Insights 9
1.5.4 Challenges and Considerations 10
1.5.5 Future Directions 10
1.6 Challenges and Future Directions 10
1.6.1 Data Quality Issues 11
1.6.2 Model Interpretability 11
1.6.3 Overfitting 11
1.6.4 Interdisciplinary Collaboration 11
1.6.5 Future Directions 12
1.7 Conclusion 12

Chapter 2 Research Methodology 13
2.1 Data Preparation 13
2.1.1 Data Sources 13
2.1.2 Data Integration 15
2.1.3 Data Cleaning and Preprocessing 19
2.2 Research Design 25
2.3 Exploratory Data Analysis (EDA) 25
2.3.1 Descriptive Statistics: 25
2.3.2 Correlation Analysis: 29
2.4 Feature Selection and Engineering 31
2.4.1 Selection of Key Variables 31
2.4.2 Feature Engineering 31
2.4.3 Integration with Analysis Objectives 33
2.5 Model Development 33
2.5.1 Algorithm Selection 33
2.5.2 Hyperparameter Tuning 34
2.6 SAS Model Studio 38
Chapter 3 Analysis and Research Findings 41
3.1 Model Evaluation Metrics: 41
Average Squared Error (ASE): 41
Observed Average: 41
Sum of Squared Errors (SSE): 41
Observations Used: 41
3.2 Gradient Boosting for Feature Selection 41
3.3 Univariate Analysis 43
3.3.1 Type Variable 43
3.3.2 Minutes Played Variable 44
3.3.3 Minute of Event Variable 45
3.3.4 National Players Variable 47
3.4 Clustering Analysis 48

3.4.1 Home and Away Goals Clustering 48
3.4.2 Assists and Goals Clustering 49
3.5 Boxplot Analysis 50
3.6 Conclusion 51
Chapter 4 Discussion 53
4.1 Comparison with Literature 53
4.1.1 Existing Literature 53
4.1.2 Current Study 53
4.2 Implications for Player Performance Analysis 53
4.2.1 Substitution Strategy Optimization 53
4.2.2 Player Development and Utilization 54
4.2.3 Tactical Adjustments 54
4.3 Potential Limitations of the Study 54
4.3.1 Data Limitations 54
4.3.2 Model Limitations 55
4.3.3 Contextual Factors 55
4.4 Suggestions for Further Research and Improvement 56
4.4.1 Integration of Real-Time Data 56
4.4.2 Advanced Analytical Techniques 56
4.4.3 Contextual and External Factors 56
4.4.4 Interdisciplinary Collaboration 56
4.4.5 Enhanced Visualization and Reporting 57
4.5 Interpretation and Strategic Implications 57
4.6 Feature Importance Analysis 58
4.7 Future Research Directions 59
4.7.1 Incorporating Real-Time Data 59
4.7.2 Exploring Advanced Machine Learning Techniques 59
4.8 Enhanced Collaboration 59
Chapter 5 Conclusion and Future prospects 60
Chapter 6 References 61

List of Figures

Figure 1.1-i: Appearances csv file

Figure 1.1-ii: Game events csv file

Figure 1.1-iii: Games csv file

Figure 1.1-iv: Clubs csv file

Figure 1.1-v: Jupyter Notebook unable to handle processing

Figure 1.1-vi: Joining appearances and game events data in SAS Viya

Figure 1.1-vii: Joining resulting data with games data in SAS Viya

Figure 1.1-viii: Joining resulting data with clubs data in SAS Viya

Figure 1.1-ix: Converting columns in SAS Viya

Figure 1.1-x :Converting columns in SAS Viya

Figure 1.1-xi: Removing columns in SAS Viya

Figure 1.1-xii: Renaming columns in SAS Viya

Figure 1.1-xiii:Changing aggregation

Figure 1.1-xiv: Changing aggregation

Figure 1.1-xv: Final Dataset

Figure 1.1-xvi: Time series plot for away goals and club position

Figure 1.1-xvii: Time series plot for home goals and club position

Figure 1.1-xviii: Total cards Histogram

Figure 1.1-xix: Correlation Matrix

Figure 1.1-xx: Heatmap for Home advantage

Figure 1.1-xxi: Total cards feature engineering on SAS Viya

Figure 1.1-xxii: Home advantage feature engineering on SAS Viya

Figure 1.1-xxiii Hyperparameter Tuning on SAS Viya

Figure 1.1-xxiv: 50 tree model

Figure 1.1-xxv: 100 tree model

Figure 1.1-xxvi: 1000 tree model

Figure 1.1-xxvii: Importance of Key features

Figure 1.1-xxviii: Data verification before using it on SAS Model Studio

Figure 1.1-xxix: Gradient Boosting Pipeline

Figure 1.1-xxx: Multi-algorithm Pipeline

Figure 1.1-xxxi: Comparison between the Pipelines and the models used

Figure 1.1-xxxii: Important features across all pipelines

Figure 1.1-xxxiii: Actual vs Predicted Mean of goals in the best model

Figure 1.1-xxxiv: Assessment of the three attempts at 50,100 and 1000 trees

Figure 1.1-xxxv : Automatic Explanation of type variable

Figure 1.1-xxxvi: Explanation description on type variable

Figure 1.1-xxxvii: Automatic Explanation of minutes played variable

Figure 1.1-xxxviii: Automatic Explanation of minute of event variable

Figure 1.1-xxxix: Automatic Explanation of National Players variable

Figure 1.1-xl: Home and Away Goals Clustering

Figure 1.1-xli: Cluster centroids for assists and goals

Figure 1.1-xlii: Boxplot of Event types distributed over minute of the event

Figure 1.1-xliii: Descriptive statistics through Boxplot

Acknowledgements

Dr. Liam Naughton for Introducing SAS Viya.

Dr. Andrew Gascoyne for clear and concise guidance as module leader.

Pooja Kaur for supporting and guiding every step of the project.

Chapter 1 Literature Review

1.1 Introduction

This literature review examines the growing field of football analytics, particularly the application of machine learning algorithms to predict player ratings and match outcomes. The review begins by exploring the evolution of football analytics, tracing its development from basic statistics to sophisticated video analysis and GPS tracking technologies. It then analyses the various match analysis systems employed to capture player movements and performance metrics. Shifting focus to machine learning, the review examines its role in predicting player ratings. It explores the types of algorithms used, the player attributes analysed, and the value these predictions bring to various stakeholders in football.

Next, the review investigates how machine learning tackles the challenge of predicting match outcomes. It discusses the variables analysed, the methodologies employed, and the inherent challenges faced by these models. Finally, the review examines the key challenges of data quality, model interpretability, and overfitting in football analytics. It concludes by discussing future directions, including advancements in data collection methodologies, the development of clearer models, and fostering collaboration between data scientists and football experts.

1.2 Evolution of Football Analytics

Historically, football analytics was primarily based on basic statistics and subjective observations, limiting the depth of insights into player performance and team strategies. Coaches and analysts relied on traditional metrics such as goals scored, assists, and possession percentage to assess player contributions and match outcomes. However, these metrics provided only a surface-level understanding of the game, often overlooking the nuanced aspects of player performance and team dynamics.

The evolution of football analytics began with advancements in technology and data collection methods. The introduction of video analysis tools allowed analysts to capture and review match footage in detail, providing a more comprehensive understanding of player movements and tactical patterns. Video-based time-motion analysis systems emerged as a valuable tool for quantifying player activities such as distance covered, sprints made, and positioning on the field. While these systems

offered more detailed insights than traditional statistics, they still relied on manual data collection and subjective interpretation. (Lawrence, Crawford 2021)

The next breakthrough in football analytics came with the development of semi-automatic multiple-camera systems and GPS technology. These systems enabled real-time tracking of player movements and interactions on the field, providing objective data on player positioning, speed, and acceleration. Semi-automatic multiple-camera systems utilized computer vision algorithms to track player movements automatically, reducing the need for manual intervention and improving data accuracy. Similarly, GPS technology allowed for continuous monitoring of player performance metrics during training sessions and matches, offering valuable insights into player fitness levels and workload management. (Sarmento, Clemente et al. 2017)

The integration of sophisticated analytical techniques, including machine learning algorithms, further revolutionized football analytics. Machine learning algorithms can analyse vast amounts of data and identifying complex patterns and relationships that may not be apparent to human analysts. In the context of football, these algorithms can predict player ratings, suggest match outcomes, and uncover hidden insights from player attributes and match data. By leveraging historical data and advanced analytical techniques, machine learning algorithms offer coaches, analysts, and stakeholders in the football industry unprecedented opportunities to optimize player evaluation, team strategies, and performance prediction.

1.3 Comparing Football Match Analysis Systems

Football match analysis systems have evolved significantly over the years, with various technologies being employed to capture and analyse player movements and performance during matches. In this section, we will explore the different types of football match analysis systems, their strengths and limitations, and the impact of system choice on the accuracy and reliability of performance metrics. (Baattite 2023)

1.3.1 Video-Based Time-Motion Analysis

Video-based time-motion analysis systems involve manually reviewing match footage to track player movements and activities on the field. Analysts annotate key events

such as passes, shots, tackles, and sprints, and then quantify player activities based on the annotated data. (Spencer, Hawkey et al.)

Strengths:

- Provides detailed insights into player movements, positioning, and interactions during matches.
- Allows for subjective analysis of player performance and tactical strategies.
- Can capture nuanced aspects of player behaviour that may not be evident from traditional statistics.

Limitations:

- Relies on manual data collection and annotation, which can be time-consuming and subjective.
- Limited by the field of view of the camera and the quality of the footage, which may impact the accuracy of the analysis.
- Difficult to scale for large datasets or real-time analysis during matches.

1.3.2 Semi-Automatic Multiple-Camera Systems

Semi-automatic multiple-camera systems utilize computer vision algorithms to automatically track player movements on the field. Multiple cameras are strategically

placed around the stadium to capture different angles and perspectives, allowing for comprehensive coverage of the match. (Spencer, Hawkey et al.)

Strengths:

- Offers a more objective and automated approach to player tracking compared to video-based systems.
- Can capture a wide range of player activities and movements from multiple angles.
- Provides real-time data during matches, enabling immediate analysis and feedback.

Limitations:

- Requires significant infrastructure and setup costs, including the installation of multiple cameras and calibration of the system.
- May still require manual intervention for data validation and error correction.
- Limited by the line of sight and field of view of the cameras, which may affect the accuracy of tracking in certain areas of the field.

1.3.3 GPS Systems

GPS systems utilize wearable tracking devices worn by players to monitor their movements, speed, and distance covered during matches. These devices use GPS

technology to collect and transmit data in real-time, providing objective insights into player performance and workload. (Spencer, Hawkey et al.)

Strengths:

- Offers objective and precise measurements of player movements and physical exertion.
- Provides real-time data during matches, allowing coaches and analysts to monitor player fitness and workload.
- Can track player movements across the entire field, regardless of camera angles or field of view limitations.

Limitations:

- Limited to outdoor matches and may be affected by environmental factors such as signal interference or satellite coverage.
- Requires players to wear additional equipment, which may impact their comfort and performance.
- May not capture certain aspects of player performance, such as technical skills or tactical awareness, that are better analysed using other systems.

1.3.4 Impact of System Choice

The choice of match analysis system significantly impacts the accuracy and reliability of performance metrics in football analytics. Each system has its strengths and limitations, and the optimal choice depends on factors such as the specific goals of the analysis, available resources, and logistical constraints. By understanding the capabilities and limitations of each system, analysts can make informed decisions to ensure the accuracy and reliability of performance metrics in football analytics.

1.4 Utilizing Machine Learning Algorithms for Player Rating Prediction

Machine learning algorithms have emerged as powerful tools for predicting player ratings in football analytics. By analysing various player attributes such as physical, technical, and mental characteristics, these algorithms offer valuable insights into player performance and contribute to informed decision-making processes for

coaches, analysts, and stakeholders in the football industry. (Goes, Meerhoff et al. 2020)

1.4.1 Role of Machine Learning Algorithms

Machine learning algorithms play a crucial role in predicting player ratings by leveraging historical data and advanced analytical techniques. These algorithms can

identify complex patterns and relationships between player attributes and performance metrics, enabling accurate assessments of player performance.

1.4.2 Analysed Player Attributes

In predicting player ratings, machine learning algorithms analyse a wide range of player attributes, including but not limited to:

- Physical attributes such as height, weight, and speed.
- Technical skills such as passing accuracy, shooting proficiency, and ball control.
- Mental characteristics such as decision-making ability, tactical awareness, and leadership qualities.

By considering these attributes collectively, machine learning algorithms can provide a comprehensive evaluation of player performance.

1.4.3 Employed Algorithms

Various machine learning algorithms have been employed to develop predictive models for player rating prediction, including:

Regression: Linear regression and logistic regression models are commonly used to predict continuous or categorical player ratings based on input features.

Decision Trees: Decision tree algorithms partition the feature space into hierarchical decision nodes to predict player ratings.

Random Forests: Random Forest algorithms aggregate predictions from multiple decision trees to improve accuracy and robustness.

Neural Networks: Deep learning techniques, such as artificial neural networks, utilize complex architectures to learn nonlinear relationships between player attributes and performance ratings.

Each algorithm has its strengths and limitations, and the choice of algorithm depends on factors such as the complexity of the data and the desired predictive accuracy.

1.4.4 Insights and Value

The predictive models developed using machine learning algorithms offer valuable insights into player performance and contribute to data-driven decision-making in football. Coaches and analysts can use these models to identify key areas for

improvement, optimize player selection and team formation, and develop targeted training programs to enhance player development.

1.4.5 Future Directions

While machine learning algorithms have shown promise in predicting player ratings, there is still room for improvement and innovation in the field. Future research could focus on refining existing algorithms, integrating additional data sources, and exploring new techniques to further enhance the accuracy and reliability of player rating predictions. Additionally, interdisciplinary collaboration between data scientists, football analysts, and coaches is essential to ensure the practical relevance and applicability of predictive models in real-world football settings. (Goes, Meerhoff et al. 2020)

1.5 Predicting Match Outcomes with Machine Learning

Machine learning algorithms have become instrumental in forecasting match outcomes in football analytics. By analysing historical match data, player attributes, and contextual factors, these algorithms offer valuable insights into the likelihood of a team winning, losing, or drawing a match. Incorporating a wide range of variables such as opponent strength, match conditions, and team dynamics, these models

provide more accurate and nuanced predictions, contributing to informed decision-making processes in the football industry. (Tenga)

1.5.1 Analysed Variables

In predicting match outcomes, machine learning algorithms analyse various variables, including:

Historical match data: Previous match results, goals scored, goals conceded, and home/away performance records.

Player attributes: Individual player ratings, form, injuries, and suspensions.

Contextual factors: Opponent strength, match conditions (e.g., weather, stadium), and team dynamics (e.g., recent performance, managerial changes). (Peña, Touchette 2012)

By considering these variables holistically, machine learning algorithms can generate comprehensive forecasts of match outcomes. (Arntzen, Hvattum 2020)

1.5.2 Methodologies Employed

Machine learning algorithms utilize different methodologies to predict match outcomes, including:

Classification: Classifying match outcomes into win, loss, or draw based on input features and historical data.

Regression: Predicting the likelihood of specific match outcomes (e.g., win probability, goal difference) using continuous output variables.

Ensemble Methods: Combining predictions from multiple models or algorithms to improve accuracy and robustness.

Time-Series Analysis: Modelling temporal dependencies in match data to capture evolving team dynamics and performance trends over time.

These methodologies allow machine learning algorithms to capture complex patterns and relationships in match data, enhancing the accuracy of outcome predictions. (Alfredo, Isa 2019)

1.5.3 Value and Insights

The predictive models developed using machine learning algorithms offer valuable insights into match outcomes, enabling stakeholders in the football industry to make

informed decisions. Coaches, analysts, and betting agencies can use these forecasts to optimize team strategies, assess opponent strengths and weaknesses, and identify potential match-fixing or irregularities.

1.5.4 Challenges and Considerations

Despite their effectiveness, machine learning algorithms face several challenges in predicting match outcomes. These include:

Data Quality: Ensuring the accuracy, completeness, and reliability of historical match data.

Overfitting: Preventing models from capturing noise or spurious correlations in the data.

Interpretability: Understanding the rationale behind model predictions and identifying actionable insights for decision-making.

Addressing these challenges requires careful data preprocessing, model validation, and interpretation of results to ensure the reliability and practical relevance of outcome predictions. (Wakelam, Steuber et al. 2022)

1.5.5 Future Directions

Future research in predicting match outcomes with machine learning algorithms could focus on:

- Incorporating new data sources such as tracking data, social media sentiment analysis, and betting odds to enhance predictive accuracy.
- Developing models that account for dynamic factors such as in-game events, tactical adjustments, and player substitutions.
- Exploring novel techniques such as deep learning and reinforcement learning to capture complex interactions and dependencies in match data.

1.6 Challenges and Future Directions

While machine learning algorithms have demonstrated considerable promise in football analytics, several challenges persist, and future research endeavours should aim to overcome these obstacles and pave the way for further advancements. Additionally, interdisciplinary collaboration between data scientists, football analysts,

and coaches is essential to maximize the utility of data-driven insights in coaching strategies and player development programs. (Wright, Carling et al. 2017)

1.6.1 Data Quality Issues

One of the primary challenges in football analytics is ensuring the quality, completeness, and reliability of the data utilized for analysis. Data may be prone to errors, inconsistencies, or missing values, which can significantly impact the accuracy and reliability of machine learning models. Addressing data quality issues requires robust data preprocessing techniques, data validation procedures, and data governance frameworks to ensure the integrity of the analytical process. (Tenga)

1.6.2 Model Interpretability

Interpreting the rationale behind machine learning model predictions remains a significant challenge in football analytics. While complex algorithms may achieve high predictive accuracy, understanding how and why these models arrive at specific decisions can be challenging. Enhancing model interpretability is crucial for gaining actionable insights from the analysis and building trust among stakeholders, including coaches, analysts, and players. Future research should focus on developing techniques for explaining model predictions, visualizing model behaviour, and extracting meaningful insights from machine learning models. (Pretto, De Caso 2022)

1.6.3 Overfitting

Overfitting, wherein a model learns noise or spurious correlations in the training data rather than generalizable patterns, is a common challenge in machine learning applications, including football analytics. Overfit models may perform well on the training data but generalize poorly to unseen data, leading to inaccurate predictions and unreliable insights. To mitigate overfitting, researchers must employ appropriate regularization techniques, cross-validation strategies, and model evaluation metrics to ensure the robustness and generalizability of machine learning models.

1.6.4 Interdisciplinary Collaboration

Effective integration of data-driven insights into coaching strategies and player development programs requires interdisciplinary collaboration between data scientists, football analysts, coaches, and players. While machine learning algorithms can provide valuable insights into player performance, team tactics, and match outcomes, translating these insights into actionable strategies requires domain expertise and contextual understanding of football dynamics. Collaborative efforts between data scientists and football practitioners can facilitate the development of

tailored analytical solutions, optimize training regimes, and inform strategic decision-making processes. (ÜNSOY 2022)

1.6.5 Future Directions

Advancing data collection methodologies to capture richer and more diverse datasets, including tracking data, physiological measurements, and biometric data. (Wakelam, Steuber et al. 2022)

- Developing interpretable machine learning models that provide transparent explanations of their predictions and recommendations.
- Exploring innovative techniques such as deep learning, reinforcement learning, and causal inference to address complex analytical challenges and uncover deeper insights from football data.
- Promoting interdisciplinary collaboration and knowledge exchange between data scientists, football analysts, coaches, and players to co-create data-driven solutions and enhance performance outcomes on and off the field.

1.7 Conclusion

Machine learning has revolutionized football analytics. It goes beyond basic stats, offering deep dives into player performance, team dynamics, and even predicting match outcomes. This data-driven approach gives coaches and analysts a competitive edge.

Advanced systems like video analysis and GPS tracking, combined with machine learning algorithms, allow for highly detailed player movement analysis and contextual understanding. Player ratings and match predictions are becoming increasingly accurate.

However, challenges like data quality and model interpretability remain. To address these, interdisciplinary collaboration and ongoing research are crucial to develop reliable and clear models.

The future of football analytics lies in better data collection, clearer models, and innovative techniques. By embracing this approach, football analytics will continue to empower stakeholders with valuable insights, ultimately optimizing performance and driving success on the pitch.

Chapter 2 Research Methodology

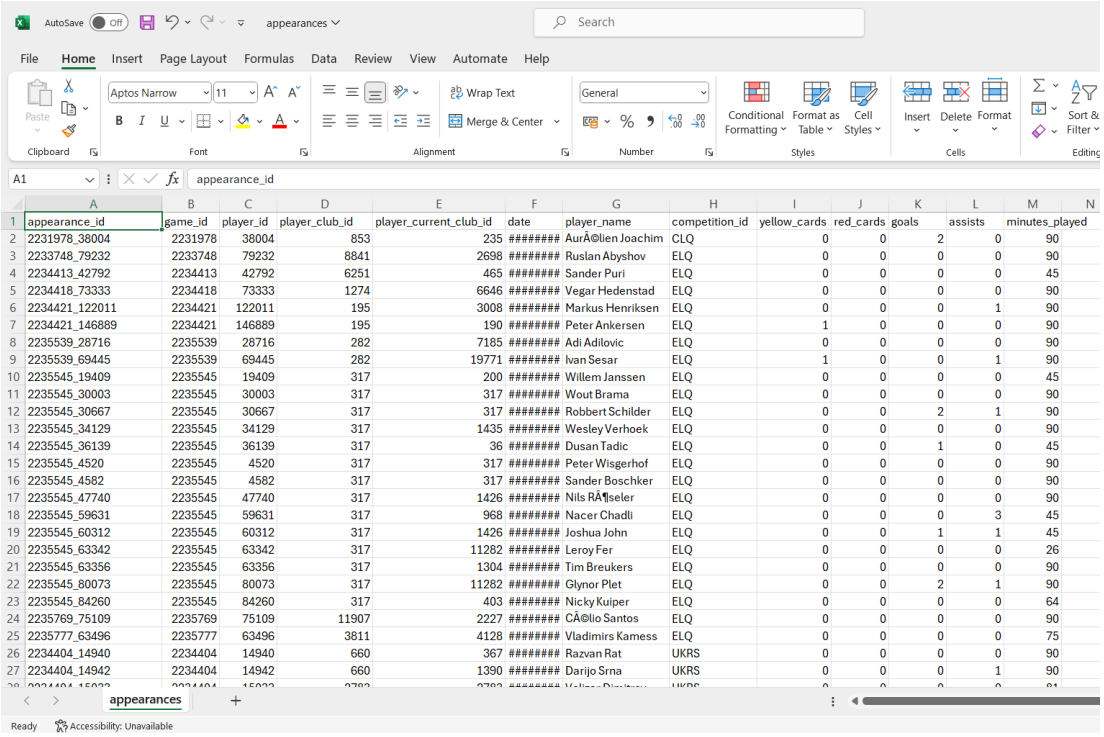
2.1 Data Preparation

The data preparation phase is fundamental in ensuring the robustness and reliability of the subsequent analysis. This section outlines the systematic approach taken to prepare the datasets for a comprehensive analysis, which aims to investigate the impact of various game events on player effectiveness and match outcomes.

2.1.1 Data Sources

This analysis integrates data from multiple sources, each offering distinct insights into football matches and player performances:

Appearances Data (appearances.csv): This dataset provides detailed information about player appearances in matches, including variables such as player_id, game_id, yellow_cards, red_cards, goals, assists, and minutes_played.

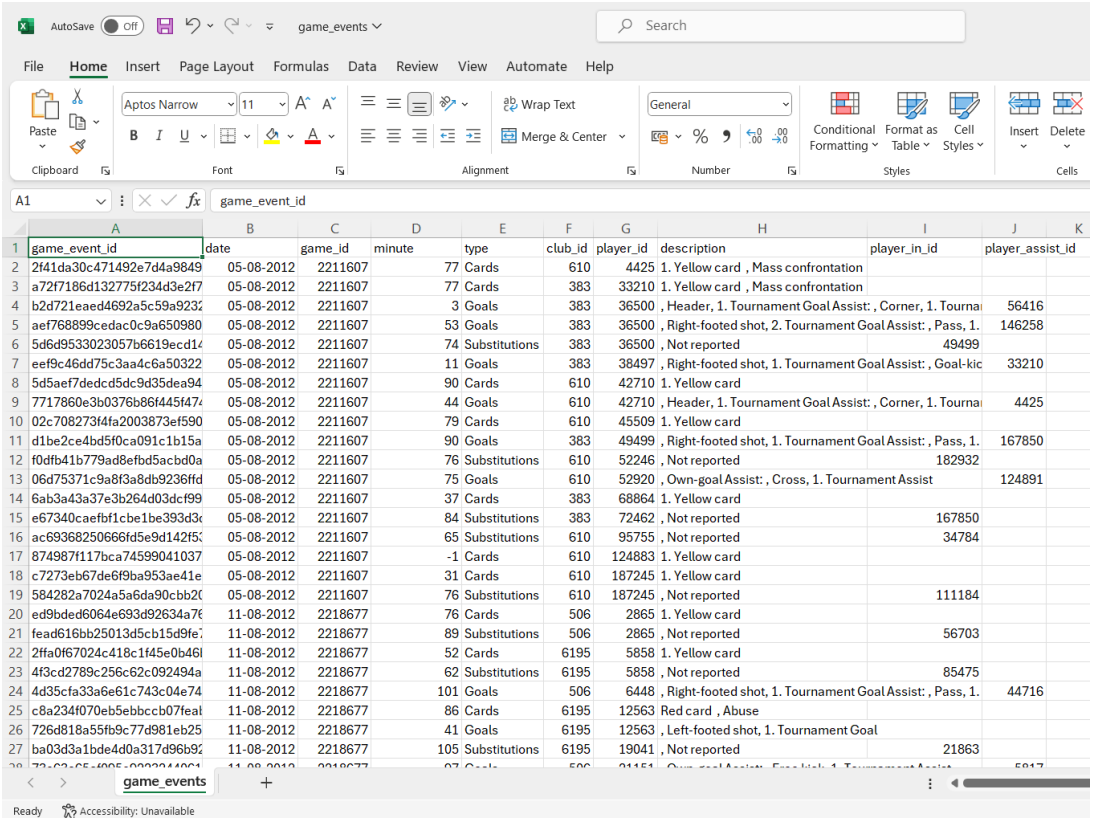


appearance_id	game_id	player_id	player_club_id	player_current_club_id	date	player_name	competition_id	yellow_cards	red_cards	goals	assists	minutes_played
2231978_38004	2231978	38004	853	235	########	AurÃ©lien Joachim	CLQ	0	0	2	0	90
2233748_79232	2233748	79232	8841	2698	########	Ruslan Abyshov	ELQ	0	0	0	0	90
2234413_42792	2234413	42792	6251	465	########	Sander Puri	ELQ	0	0	0	0	45
2234418_73333	2234418	73333	1274	6646	########	Vegar Hedenstad	ELQ	0	0	0	0	90
2234421_122011	2234421	122011	195	3008	########	Markus Henriksen	ELQ	0	0	0	1	90
2234421_146889	2234421	146889	195	190	########	Peter Ankersen	ELQ	1	0	0	0	90
2235539_28716	2235539	28716	282	7185	########	Adi Adilovic	ELQ	0	0	0	0	90
2235539_69445	2235539	69445	282	19771	########	Ivan Sesar	ELQ	1	0	0	1	90
2235545_19409	2235545	19409	317	200	########	Willem Janssen	ELQ	0	0	0	0	45
2235545_30003	2235545	30003	317	317	########	Wout Brama	ELQ	0	0	0	0	90
2235545_30667	2235545	30667	317	317	########	Robbert Schilder	ELQ	0	0	2	1	90
2235545_34129	2235545	34129	317	1435	########	Wesley Verhoek	ELQ	0	0	0	0	90
2235545_36139	2235545	36139	317	36	########	Dusan Tadic	ELQ	0	0	1	0	45
2235545_4520	2235545	4520	317	317	########	Peter Wisgerhof	ELQ	0	0	0	0	90
2235545_4582	2235545	4582	317	317	########	Sander Boschker	ELQ	0	0	0	0	90
2235545_47740	2235545	47740	317	1426	########	Nils RÃ¶seler	ELQ	0	0	0	0	90
2235545_59631	2235545	59631	317	968	########	Nacer Chadli	ELQ	0	0	0	3	45
2235545_60312	2235545	60312	317	1426	########	Joshua John	ELQ	0	0	1	1	45
2235545_63342	2235545	63342	317	11282	########	Leroy Fer	ELQ	0	0	0	0	26
2235545_63356	2235545	63356	317	1304	########	Tim Breukers	ELQ	0	0	0	0	90
2235545_80073	2235545	80073	317	11282	########	Glynor Plet	ELQ	0	0	2	1	90
2235545_84260	2235545	84260	317	403	########	Nicky Kuiper	ELQ	0	0	0	0	64
2235769_75109	2235769	75109	11907	2227	########	CÃ©lio Santos	ELQ	0	0	0	0	90
2235777_63496	2235777	63496	3811	4128	########	Vladimirs Kamess	ELQ	0	0	0	0	75
2234404_14940	2234404	14940	660	367	########	Razvan Rat	UKRS	0	0	0	0	90
2234404_14942	2234404	14942	660	1390	########	Darijo Srna	UKRS	0	0	0	1	90

Figure 2.1-i: Appearances csv file

Game Events Data (game_events.csv): This dataset captures specific events during matches, with variables such as game_id, player_id, minute, type, description,

player_in_id, and player_assist_id. The ‘type ‘variable denotes the nature of the event, such as goals, substitutions, and disciplinary actions.



game_event_id	date	game_id	minute	type	club_id	player_id	description	player_in_id	player_assist_id
2f41da30c471492e7d4a9849	05-08-2012	2211607	77	Cards	610	4425	1. Yellow card , Mass confrontation		
a72f7186d132775f234d3e2f7	05-08-2012	2211607	77	Cards	383	33210	1. Yellow card , Mass confrontation		
b2d721eaed4692a5c59a9233	05-08-2012	2211607	3	Goals	383	36500	, Header, 1. Tournament Goal Assist: , Corner, 1. Tourna	56416	
aef768899cedac0c9a650980	05-08-2012	2211607	53	Goals	383	36500	, Right-footed shot, 2. Tournament Goal Assist: , Pass, 1.	146258	
5d6d9533023057b6619ecd14	05-08-2012	2211607	74	Substitutions	383	36500	, Not reported	49499	
eef9c46dd75c3aa4c6a50322	05-08-2012	2211607	11	Goals	383	38497	, Right-footed shot, 1. Tournament Goal Assist: , Goal-kic	33210	
5d5aef7dedcd5dc9d35dea94	05-08-2012	2211607	90	Cards	610	42710	1. Yellow card		
7717860e3b0376b86f445f47	05-08-2012	2211607	44	Goals	610	42710	, Header, 1. Tournament Goal Assist: , Corner, 1. Tourna	4425	
02c708273f4fa2003873ef590	05-08-2012	2211607	79	Cards	610	45509	1. Yellow card		
d1be2ce4bd5f0ca091c1b15a	05-08-2012	2211607	90	Goals	383	49499	, Right-footed shot, 1. Tournament Goal Assist: , Pass, 1.	167850	
f0dfb41b779ad8efbd5acbd0a	05-08-2012	2211607	76	Substitutions	610	52246	, Not reported	182932	
06d75371c9a8f3a8db9236ffd	05-08-2012	2211607	75	Goals	610	52920	, Own-goal Assist: , Cross, 1. Tournament Assist		124891
6ab3a43a37e3b264d03dcf99	05-08-2012	2211607	37	Cards	383	68864	1. Yellow card		
e67340caefbf1cbe1be393d3	05-08-2012	2211607	84	Substitutions	383	72462	, Not reported	167850	
ac69368250666fd5e9d142f5	05-08-2012	2211607	65	Substitutions	610	95755	, Not reported	34784	
874987f117bca74599041037	05-08-2012	2211607	-1	Cards	610	124883	1. Yellow card		
c7273eb67de6f9ba953ae41e	05-08-2012	2211607	31	Cards	610	187245	1. Yellow card		
584282a7024a5a6da90cbb2	05-08-2012	2211607	76	Substitutions	610	187245	, Not reported	111184	
ed9bded6064e693d92634a7	11-08-2012	2218677	76	Cards	506	2865	1. Yellow card		
fead616bb25013d5cb15d9fe	11-08-2012	2218677	89	Substitutions	506	2865	, Not reported	56703	
2ffa0f67024c418c1f45e0b46	11-08-2012	2218677	52	Cards	6195	5858	1. Yellow card		
4f3cd2789c256c62c092494a	11-08-2012	2218677	62	Substitutions	6195	5858	, Not reported	85475	
4d35cfa33a6e61c743c04e74	11-08-2012	2218677	101	Goals	506	6448	, Right-footed shot, 1. Tournament Goal Assist: , Pass, 1.	44716	
c8a234f070eb5ebbccb07fea	11-08-2012	2218677	86	Cards	6195	12563	Red card , Abuse		
726d818a55fb9c77d981eb25	11-08-2012	2218677	41	Goals	6195	12563	, Left-footed shot, 1. Tournament Goal		
ba03d3a1bde4d0a317d96b9	11-08-2012	2218677	105	Substitutions	6195	19041	, Not reported	21863	

Figure 2.1-ii: Game events csv file

Games Data (games.csv): This dataset contains comprehensive match details, including game_id, competition_id, season, round, date, home_club_id,

away_club_id, home_club_goals, away_club_goals, home_club_name, and away_club_name.

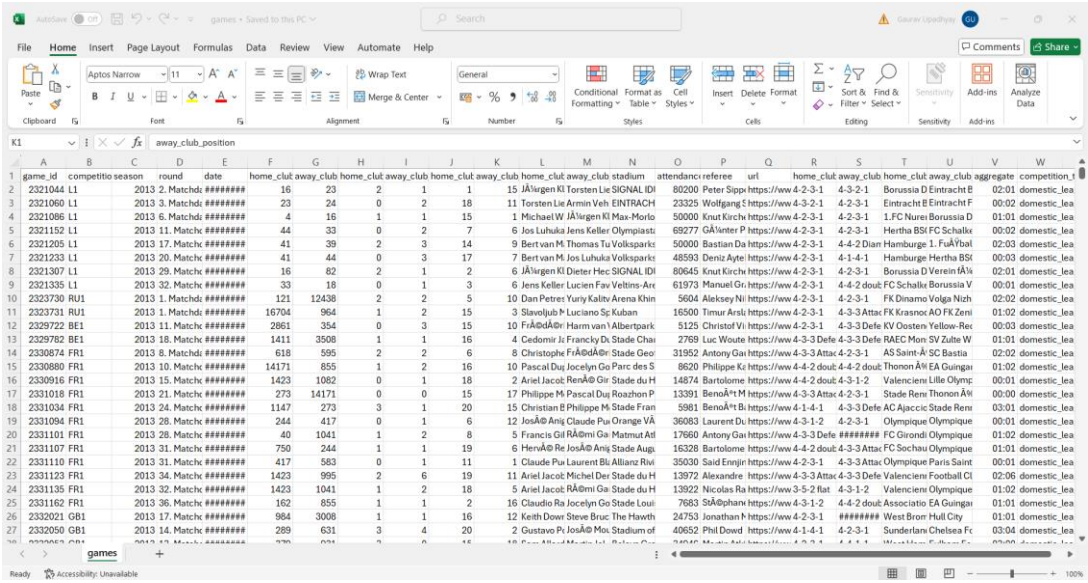


Figure 2.1-iii: Games csv file

Club Information Data (clubs.csv): This dataset provides information about football clubs, including club_id, name, total_market_value, squad_size, average_age, foreigners_number, foreigners_percentage, and national_team_players.

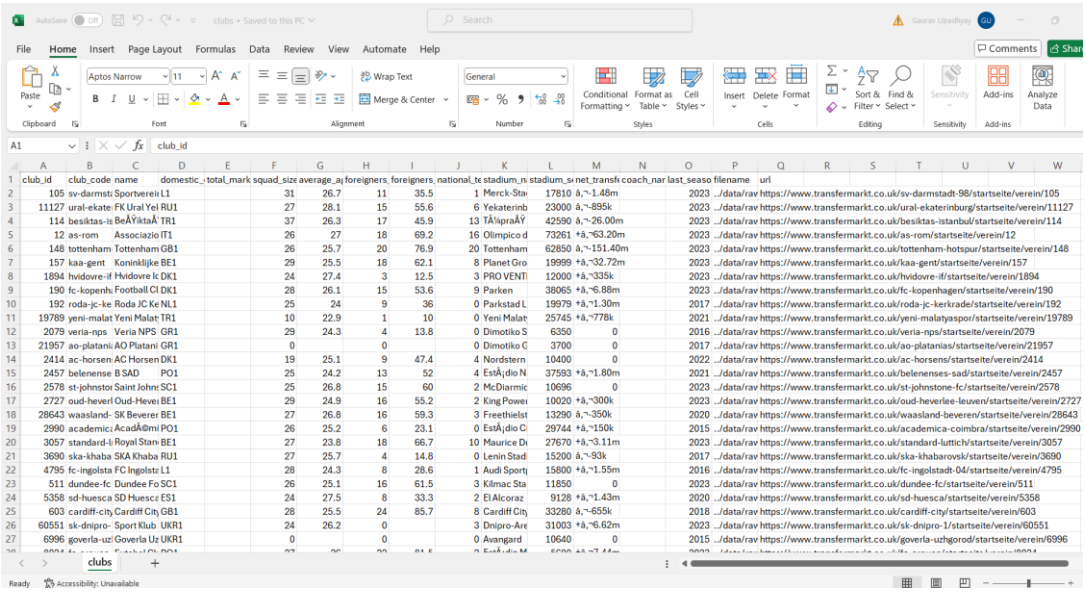


Figure 2.1-iv: Clubs csv file

2.1.2 Data Integration

Merging Appearances and Game Events Data

```
    'players': 'C:/Users/91961/Dropbox/My PC (LAPTOP-7M5BBVVO)/Downloads/archive (1)/players.csv'
}

# Read and optimize the initial DataFrame
combined_df = pd.read_csv(csv_files['appearances'])
combined_df = optimize_dataframe(combined_df)
print(f"Initial DataFrame shape: {combined_df.shape}, memory usage: {combined_df.memory_usage(deep=True).sum() / 1024**2:.2f} MB"

# Function to merge DataFrames in chunks
def merge_in_chunks(left_df, right_df, on, how='inner', chunk_size=50000):
    chunks = []
    for i in range(0, len(right_df), chunk_size):
        chunk = right_df.iloc[i:i + chunk_size]
        chunk = optimize_dataframe(chunk)
        merged_chunk = left_df.merge(chunk, on=on, how=how)
        chunks.append(merged_chunk)
    return pd.concat(chunks, ignore_index=True)

# Perform inner joins iteratively in chunks
merge_columns = {
    'club_games': 'game_id',
    'games': 'game_id',
    'player_valuations': 'player_id',
    'players': 'player_id',
    'clubs': 'club_id',
    'competitions': 'competition_id',
    'game_events': ['game_id', 'player_id'],
    'game_lineups': ['game_id', 'player_id']
}

for key, on in merge_columns.items():
    right_df = pd.read_csv(csv_files[key])
    print(f"Merging with {key}, DataFrame shape: {right_df.shape}, memory usage: {right_df.memory_usage(deep=True).sum() / 1024**
    combined_df = merge_in_chunks(combined_df, right_df, on=on)
    print(f"Shape after merging with {key}: {combined_df.shape}, memory usage: {combined_df.memory_usage(deep=True).sum() / 1024*

# Save the resulting DataFrame to a new CSV file
output_file = 'combined_inner_join.csv'
combined_df.to_csv(output_file, index=False)

print("CSV files combined successfully!")

# Create a download link for the CSV file
FileLink(output_file)
```

Figure 2.1-v: Jupyter Notebook unable to handle processing

The objective of this research was to merge player appearance records with game events to provide a detailed view of player actions and outcomes. Initially, eight separate CSV files containing various aspects of the data were available. The initial attempt to merge all eight files into a single comprehensive dataset using Python Jupyter Notebook and SAS Viya encountered significant challenges. These difficulties

arose due to the data's complexity and inconsistencies, making it impractical to combine all files seamlessly.

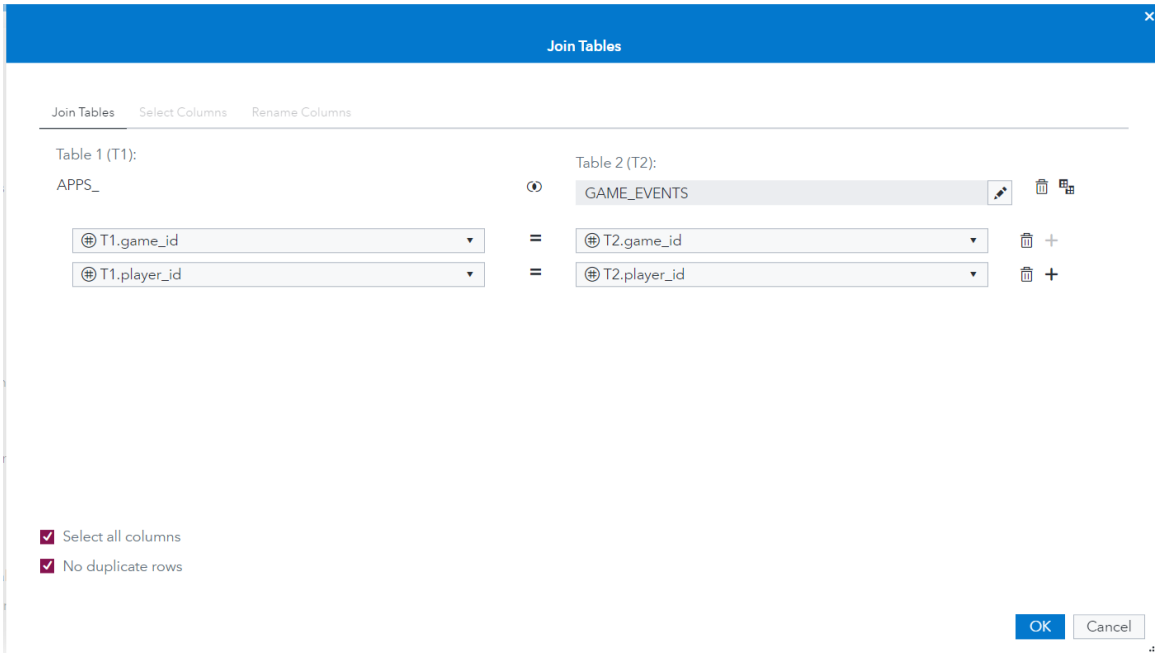


Figure 2.1-vi: Joining appearances and game events data in SAS Viya

To overcome these challenges, the research team concluded that the analysis could be effectively conducted using a subset of the data. Consequently, three to four key CSV files, which contained the most relevant information for the study, were selected. This strategic selection aimed to simplify the data integration process while preserving the essential details needed for the analysis.

The chosen files were merged by focusing on the appearances data and the game events data, utilizing game_id and player_id as common keys. This approach facilitated the creation of a dataset that captured both player appearances and their corresponding events during matches. The merged dataset, provided a comprehensive view of each player's actions and outcomes in the context of the games they participated in.

By narrowing the scope and focusing on the most pertinent data, the research effectively addressed the initial challenges and succeeded in creating a detailed dataset. This streamlined approach ensured that the analysis was both manageable

and insightful, enabling a thorough examination of player performances and game events.

Incorporating Game Information

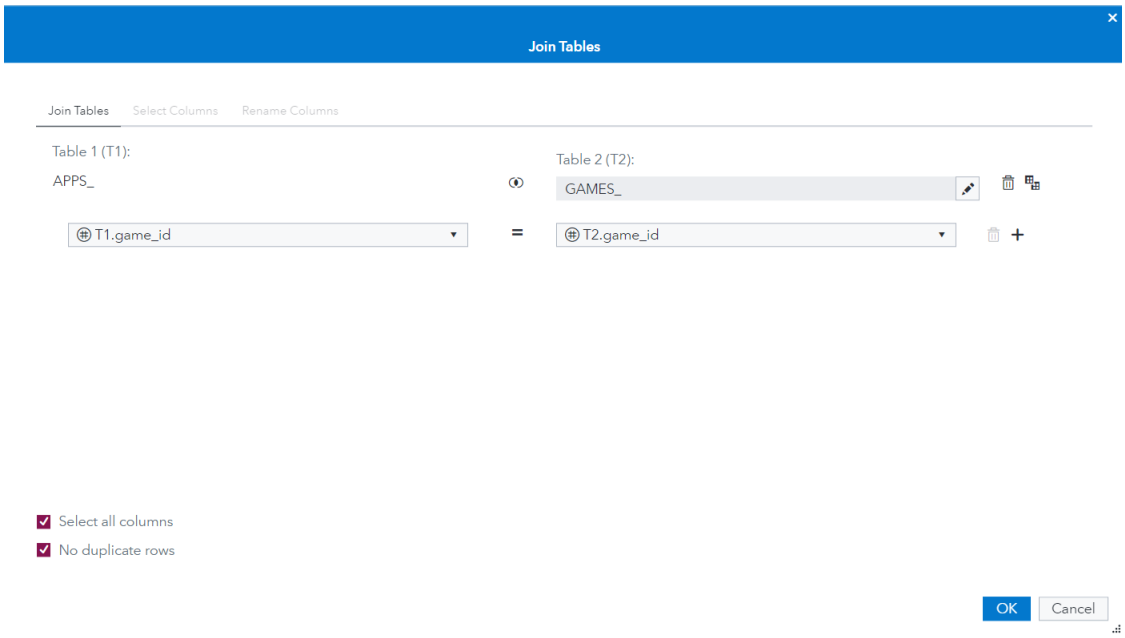


Figure 2.1-vii: Joining resulting data with games data in SAS Viya

The objective of this step was to enrich the merged dataset with additional contextual information about each match, which is crucial for understanding player actions within the specific circumstances of each game. To achieve this, the dataset—already combining player appearances and game events—was further joined with the games data using game_id as the common key. The games CSV was selected due to its comprehensive details on each match, including the date, participating clubs, and match outcomes. These details were essential for providing a complete picture of the context in which player actions occurred. By integrating this information, the dataset

was enriched with key contextual elements such as the timing of events and the competitive environment, thereby enhancing the depth and accuracy of the analysis.

Adding Club Information

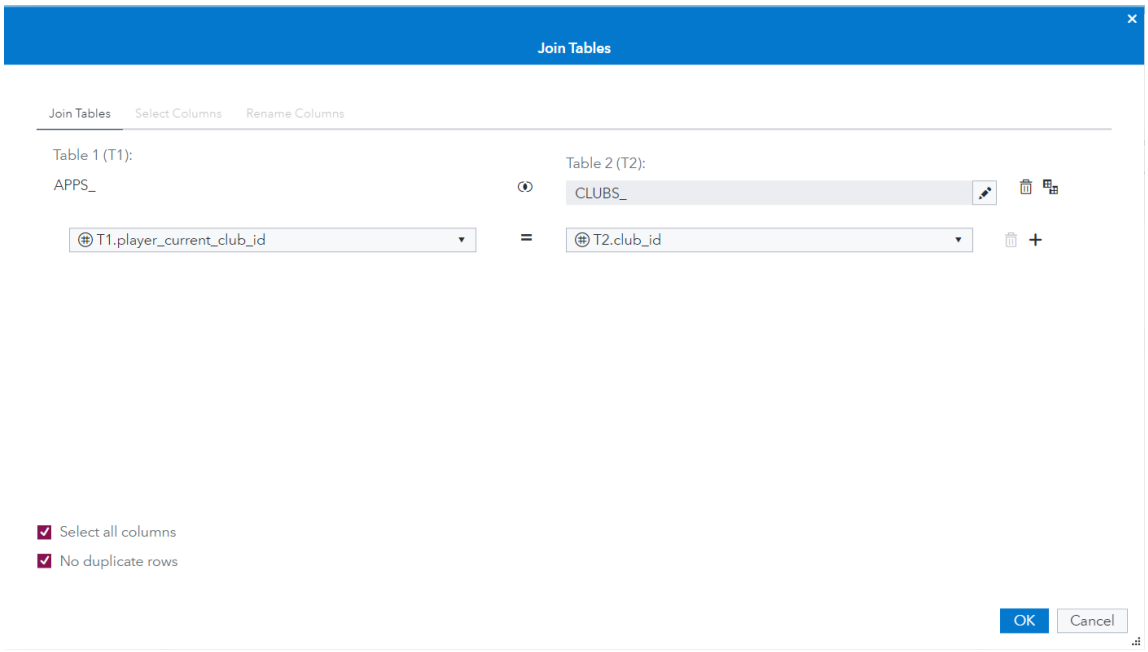


Figure 2.1-viii: Joining resulting data with clubs data in SAS Viya

The objective of this step was to further enhance the dataset by integrating club attributes, which provide broader contextual information about the environments in which the players operated. This step is vital for understanding the influence of club characteristics on player performances and match outcomes. The enriched dataset was merged with the club information data by matching home_club_id and away_club_id with club_id. The club information CSV was specifically chosen for its detailed attributes, including market value, squad size, and other club-specific characteristics. These attributes are critical for analyzing how the financial and structural aspects of clubs' impact game dynamics. By incorporating this data, the dataset was further enriched, allowing for a more comprehensive analysis that considers both player-level and club-level factors. This integration provided a holistic view, enabling a nuanced understanding of the interplay between individual performances and club contexts.

2.1.3 Data Cleaning and Preprocessing

Converting Column Data Types

The objective of this step was to ensure that all data types were correctly assigned to facilitate accurate analysis and computation. An issue was identified where certain

numeric variables were automatically converted to strings upon uploading the CSV files to SAS Viya, potentially impacting data analysis and calculations.

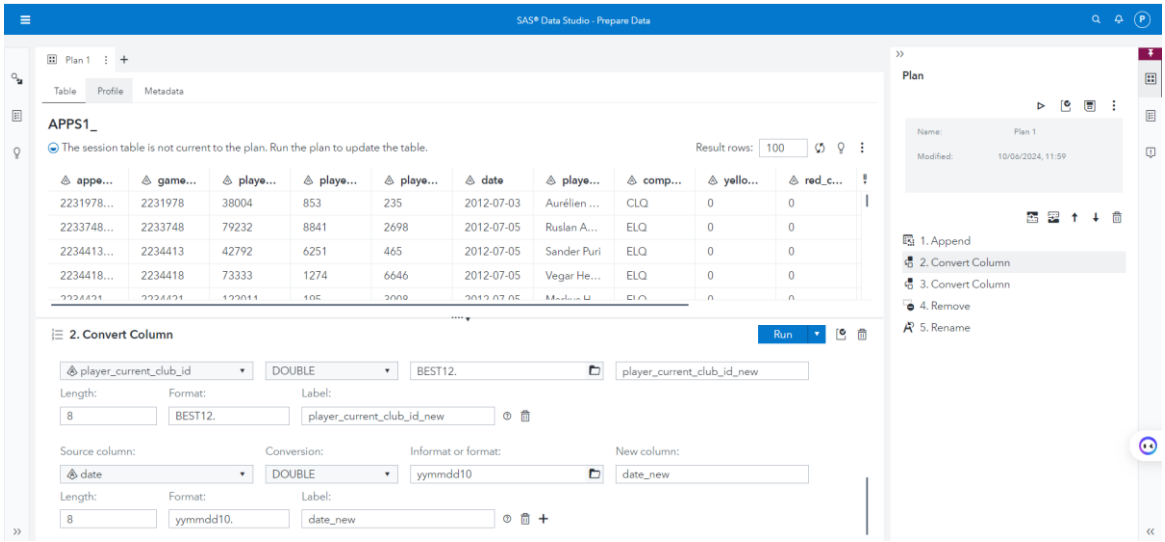


Figure 2.1-ix: Converting columns in SAS Viya

To address this, columns where data types were incorrectly assigned were identified. Key variables affected included goals, assists, minutes_played, home_club_goals, and away_club_goals. Using SAS Viya's data type conversion tools, these string variables were converted back to their appropriate numeric data types. This involved procedures to explicitly define the correct data types, ensuring that all numerical operations could be performed accurately.

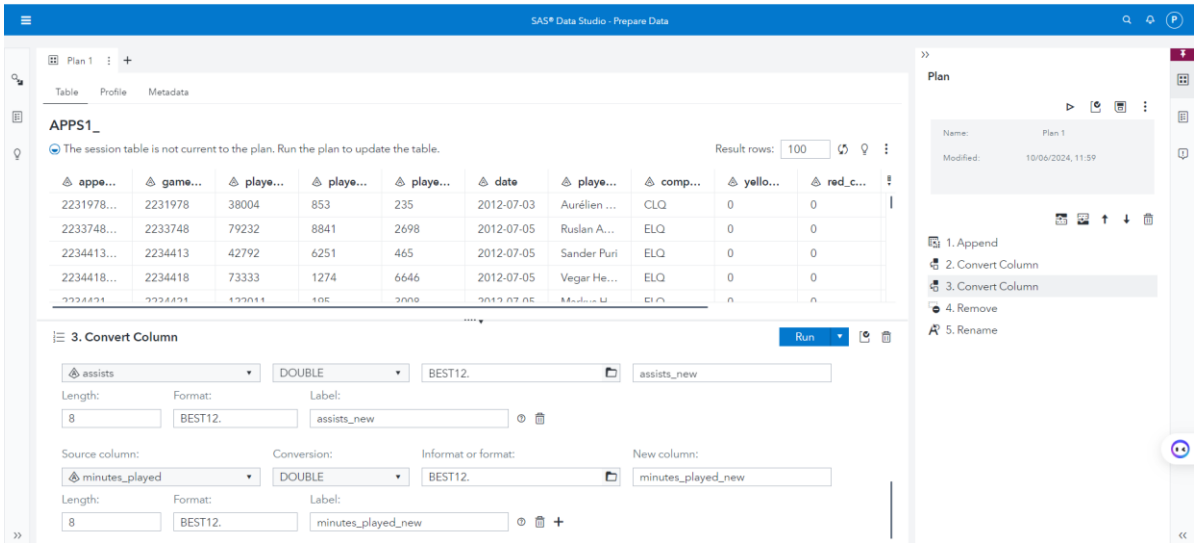


Figure 2.1-x :Converting columns in SAS Viya

Verification of the conversion was conducted by checking the summary statistics of these columns to confirm that numeric properties were restored. This step was crucial

for maintaining data integrity and ensuring that subsequent analyses were based on correctly formatted data, thereby supporting accurate and reliable results.

Removing Unnecessary Columns

The objective of this step was to streamline the dataset by removing columns that did not contribute to the analysis, thereby enhancing the focus and efficiency of the dataset. The process began with an initial review of all columns in the merged dataset to identify those that were redundant or irrelevant for the analysis.

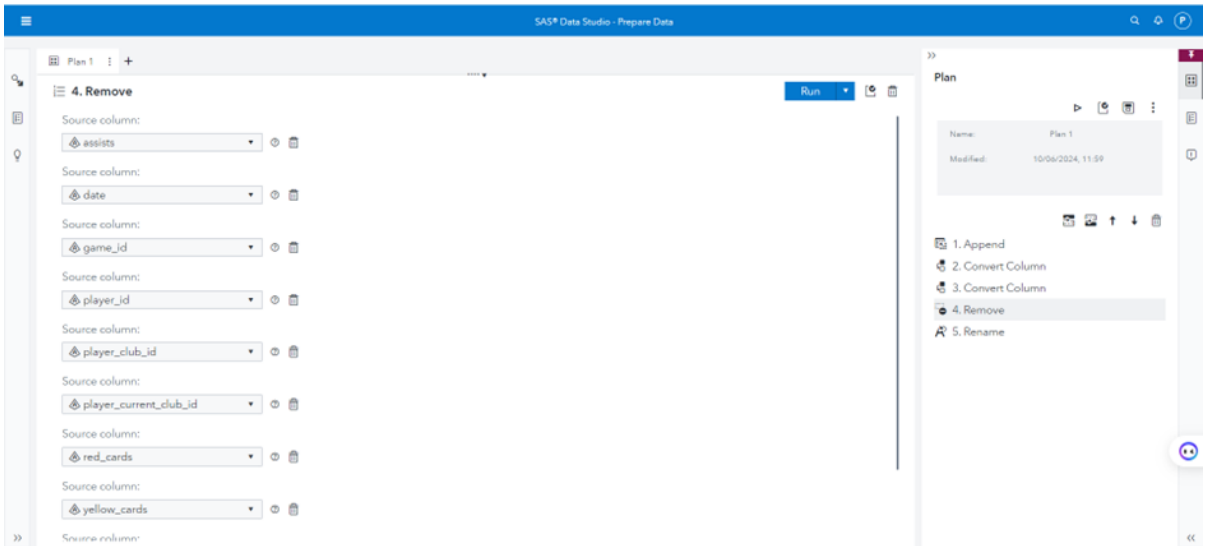


Figure 2.1-xi: Removing columns in SAS Viya

Columns such as player_in_id and player_assist_id from the game_events dataset were identified as not providing additional value to the analysis. These columns were subsequently removed using SAS Viya's data management features. The tools were used to drop these columns systematically, ensuring that the dataset was streamlined and focused only on relevant information. This step was essential for maintaining a

clean and efficient dataset, reducing complexity, and improving the overall quality of the analysis by eliminating unnecessary data. (Rønningen 2021)

Renaming Columns

The objective of this step was to standardize column names across the dataset to enhance clarity and consistency, thereby improving readability and reducing the potential for confusion during analysis.

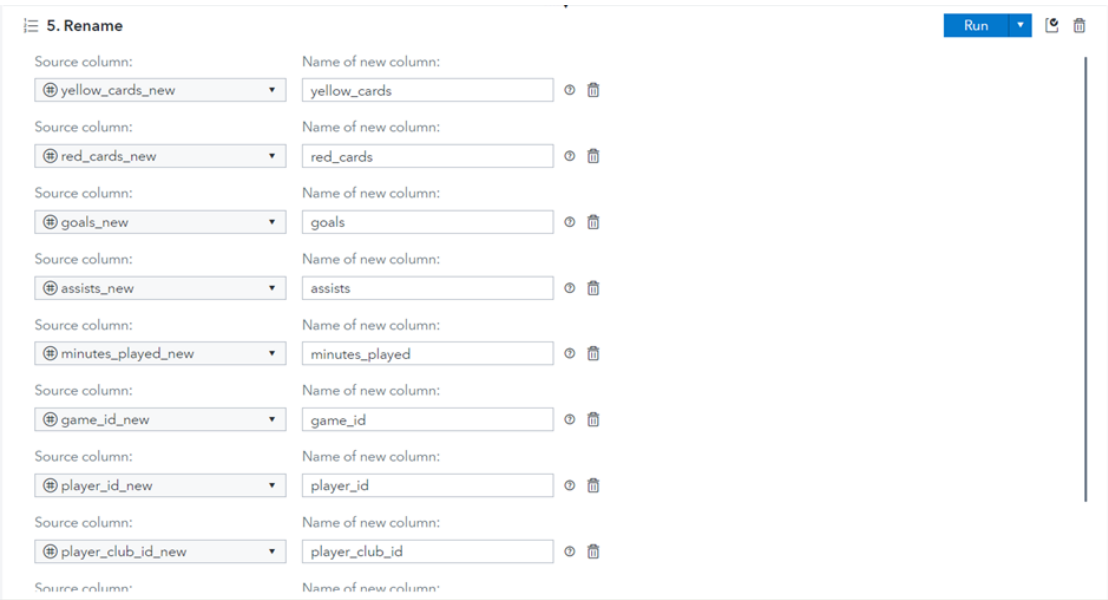


Figure 2.1-xii: Renaming columns in SAS Viya

The process began with a thorough review of all column names to ensure they were both descriptive and consistent across different datasets. Several columns were renamed to enhance readability and maintain a consistent naming convention. For instance, home_club_goals and away_club_goals were renamed to home_goals and away_goals respectively for brevity. Similarly, total_market_value was renamed to market_value to maintain a concise and clear naming convention.

These changes were implemented using SAS Viya's renaming functions, which allowed for systematic and efficient updating of column names. Additionally, all references to these columns in subsequent analyses were updated accordingly to ensure consistency throughout the dataset.

By utilizing SAS Viya's robust data management tools, the dataset was meticulously prepared, ensuring it was clean, integrated, and ready for detailed analysis. This thorough preprocessing stage laid the foundation for accurate and reliable analytical

outcomes, facilitating a smoother and more efficient analysis process. (Constantinou, Fenton 2017)

Variable Selection

The objective of this step was to enhance the usability of key variables for analysis by changing their aggregation method from 'sum' to 'average'. This subtle yet impactful adjustment was made using SAS Viya's robust data management tools, which facilitated the transformation seamlessly. Certain variables, such as minute, minutes played, and others, initially aggregated by summing their values, were recalibrated to use the average instead. This change was driven by the need for more meaningful and interpretable metrics that better reflect the nature of the data. For instance, averaging "minutes played" provides a clearer picture of player activity over multiple games rather than simply summing the total minutes.

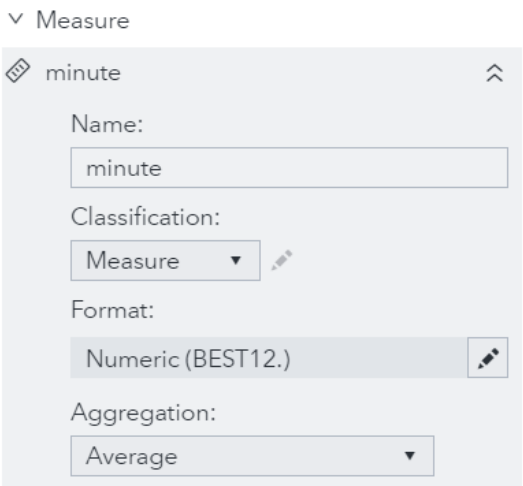


Figure 2.1-xiii:Changing aggregation

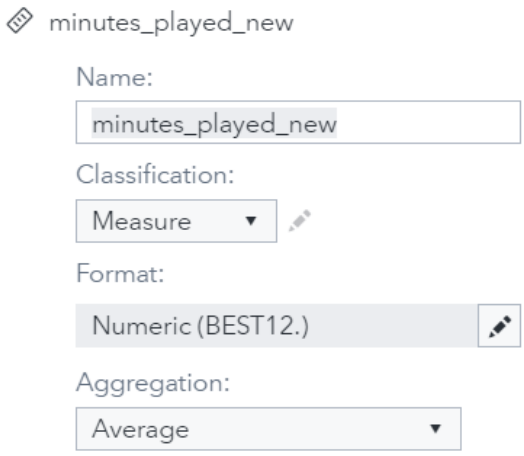


Figure 2.1-xiv: Changing aggregation

By utilizing SAS Viya's advanced aggregation functions, these variables were recalculated to reflect their average values accurately. This adjustment not only improved the clarity of the data but also ensured that the variables were more suitable for detailed analysis and modelling. This small but significant methodological improvement highlights the flexibility and power of SAS Viya in data preparation. By changing the aggregation method, the dataset became more insightful and usable for subsequent analyses, ensuring that the interpretations drawn from the data are both accurate and meaningful. This enhancement laid a stronger foundation for advanced

analytical techniques, facilitating more precise and reliable conclusions. (Herberger, Litke 2021)

Final Dataset Preparation

The objective of this step was to prepare a comprehensive and clean dataset suitable for detailed analysis. The final dataset, having undergone thorough cleaning and enrichment with relevant features, was meticulously prepared for further analysis.

aggregate	appearance_id	away_club_formation	away_club_manager_name	club_code	competition_id	competition_id_1	competition_type	date_1	domestic_competition_id	game_event_id
0:02:00	2494494_181380	4-2-3-1	Vítor Pereira	besiktas-istanbul	GR1	GR1	domestic_league	2015/02/01	TR1	00000ec947fe589
2:02:00	2982779_98968	3-5-2	Valeriy Karpin	pfk-tambov	RU1	RU1	domestic_league	2018/03/17	RU1	000054473c6c611
1:00:00	2627945_52915	4-2-3-1	Vasyl Sachko	shakhtar-donetsk	UKR1	UKR	domestic_league	2016/04/24	UKR	000077ba23a20e
0:00:00	2899887_142021	4-4-2 double 6	Abelardo	rcd-mallorca	ES1	ES1	domestic_league	2018/04/01	ES1	0000801bf49a12
1:00:00	3839818_270411	4-1-4-1	Lucien Favre	clermont-foot-63	FR1	FR1	domestic_league	2022/08/21	FR1	00009503911ab1
1:01:00	2717382_203507	4-4-2 double 6	Sérgio Conceição	sco-angers	FR1	FR1	domestic_league	2017/01/28	FR1	00009ec4f282cb2
0:01:00	3860242_291750	4-2-3-1	Ilhan Palut	konyaspor	TR1	TR1	domestic_league	2022/09/03	TR1	0000c8289784b2
5:02:00	4120719_551695	4-4-2 double 6	Yuriy Vernydub	shakhtar-donetsk	UKR1	UKR	domestic_league	2024/03/03	UKR	0001072ed2d9e8
2:02:00	3589563_274862	3-5-2	Antoine Kombouaré	fc-nantes	FR1	FR1	domestic_league	2022/04/30	FR1	0001482a9b24de
1:01:00	3886564_58884	4-4-2	Diego Simeone	real-madrid	ES1	ES1	domestic_league	2023/02/25	ES1	00016f8f3ecf71e6
1:02:00	4098040_728864	3-4-1-2	Michael Dingsdag	az-alkmaar	NL1	NL1	domestic_league	2023/12/16	NL1	000179169589b4
0:02:00	2335768_75615	4-2-3-1	Mircea Lucescu	juventus-turin	UKR1	UKR	domestic_league	2013/11/09	IT1	0001b192556e96
1:02:00	3058527_20081	4-3-1-2	Bruno Labbadia	eintracht-frankfurt	L1	L1	domestic_league	2018/12/02	L1	0001e21600ae68
3:00:00	4243701_346567	4-3-3 Attacking	Maurizio Sarri	lazio-rom	CL	CL	international_c	2024/03/05	IT1	0001fa8969885a4
3:04:00	3216235_23553	4-3-3 Attacking	Dick Advocaat	pec-zwolle	NL1	NL1	domestic_league	2020/02/16	NL1	0001fdf794a3305
1:01:00	3603131_62170	4-3-3 Attacking	Tam Courts	dundee-united-fc	SC1	SC1	domestic_league	2021/09/26	SC1	00022afb0d60c42
0:02:00	4150547_143852	4-2-3-1	Milan Rastavac	ofi-kreta	GR1	GR1	domestic_league	2023/12/21	GR1	00027b48cfb7bf
5:01:00	2883931_216603	4-4-2 double 6	Igor Tudor	fenerbahce-istanbul	TR1	TR1	domestic_league	2017/11/18	TR1	0002d18b5d4
1:03:00	3433249_286297	3-4-2-1	Paulo Fonseca	as-rom	IT1	IT1	domestic_league	2020/11/08	IT1	0002e769fd74d

Figure 2.1-xv: Final Dataset

This comprehensive dataset now includes all necessary contextual and performance-related information, enabling a detailed multivariate analysis on the 'type' variable. The preparation process ensured that all data points were accurately represented and consistent, facilitating an in-depth exploration of the relationships within the dataset.

The forthcoming analysis will focus on understanding how different types of events impact player effectiveness and match outcomes. By leveraging this enriched dataset, the analysis aims to uncover significant insights into the influence of various event types on overall performance and game dynamics, providing a robust foundation for strategic decision-making and performance evaluation. (Tuyls, Omidshafiei et al. 2021)

Conclusion

By meticulously preparing the data through these steps, the subsequent analysis using Gradient Boosting and other machine learning techniques in SAS Viya will yield robust and insightful results. Each stage of the data preparation process—from merging and cleaning to enriching and standardizing—was carefully executed to ensure data integrity and relevance.

This comprehensive data preparation process is fundamental to achieving the research objectives and deriving meaningful conclusions from the data. By ensuring that the dataset is clean, integrated, and focused on the most impactful variables, the groundwork is laid for advanced analytical techniques to uncover significant patterns and insights. This thorough approach guarantees that the analysis will be both reliable and informative, supporting accurate decision-making and meaningful outcomes. (Beal, Middleton et al. 2021)

2.2 Research Design

The research design is structured to integrate data from multiple sources, apply machine learning algorithms, and validate the models to ensure robustness and reliability. The primary objective is to predict the influence of various game events on player performance metrics and match outcomes. As detailed in the Data Preparation section, data from appearances.csv, game_events.csv, games.csv, and clubs.csv were integrated into a unified dataset. Key preprocessing steps included handling missing values, detecting and treating outliers, and feature engineering. These steps ensured the dataset was clean, comprehensive, and suitable for analysis.

2.3 Exploratory Data Analysis (EDA)

2.3.1 Descriptive Statistics:

The objective of this analysis was to explore the relationship between goal events, team rankings, and card events in football matches. Using descriptive statistics and

visualizations, the study aimed to identify patterns and critical moments that significantly influence match outcomes.

Away Goals and Team Rankings

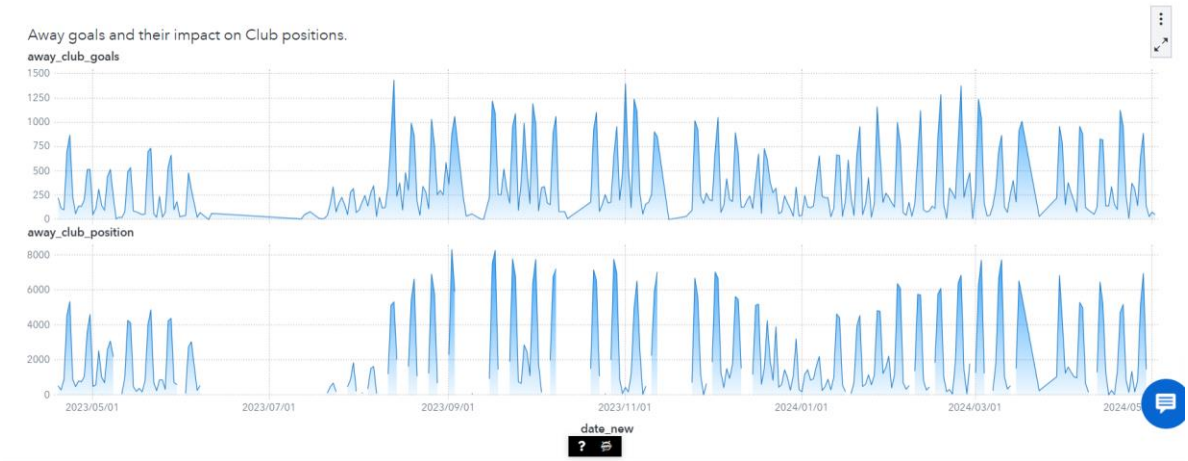


Figure 2.3-i: Time series plot for away goals and club position

The first descriptive statistic focused on away goals and their impact on team rankings, visualized through a time series plot with a shared X-axis. The analysis revealed a notable pattern: as away team goals peaked, a corresponding decrease in their rankings was observed. This inverse relationship indicates that scoring away goals is a crucial factor in improving a team's standing. Peaks in away goals often aligned with pivotal match moments, underscoring the importance of offensive strategies in away games. Teams that managed to score more away goals tended to see an immediate

improvement in their rankings, highlighting the competitive advantage gained through effective away performance.

Home Goals and Team Rankings

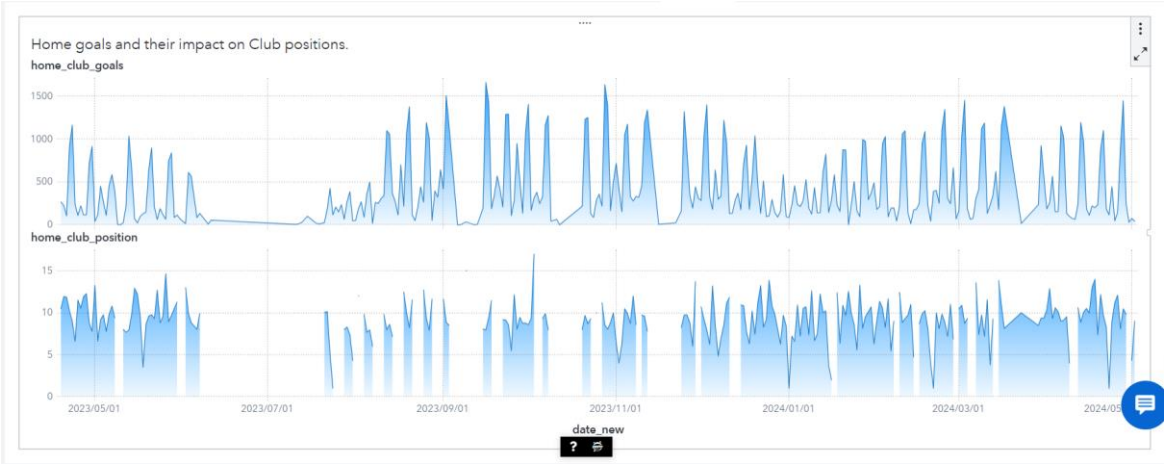


Figure 2.3-ii: Time series plot for home goals and club position

The second analysis mirrored the approach taken with away goals, this time focusing on home goals. The time series plot for home goals also shared a common X-axis to facilitate comparison. Like the away goals, an increase in home goals was often followed by a decrease in team rankings. This suggests that scoring at home plays a vital role in maintaining or enhancing a team's position. However, the peaks in home goals showed a more consistent and sustained impact on rankings compared to away goals, indicating that home advantage significantly boosts a team's performance. The

ability to capitalize on home ground conditions and convert opportunities into goals is crucial for sustaining high rankings in the league.

Timing of Card Events

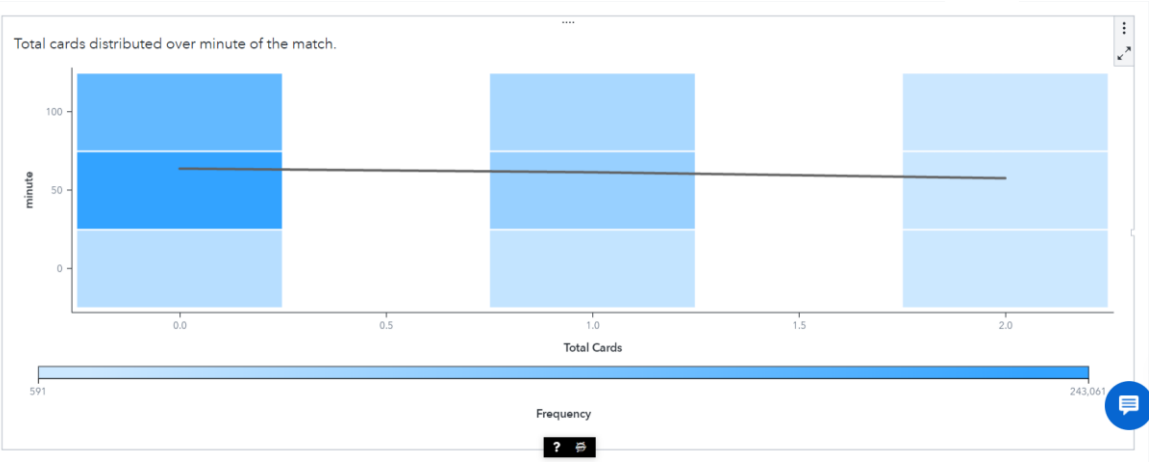


Figure 2.3-iii: Total cards Histogram

The third descriptive statistic examined the distribution of total cards over the minutes of a match, presented through a histogram. The analysis aimed to identify the most common minutes when cards were issued, providing insights into the timing and impact of disciplinary actions. The histogram revealed that cards were most frequently issued at specific intervals, particularly in the middle and towards the end of each half. This pattern suggests heightened player aggression and strategic fouls during critical phases of the match. Understanding these timings allows teams to anticipate and

manage these high-risk periods better, potentially mitigating the impact of disciplinary actions on their performance.

2.3.2 Correlation Analysis:

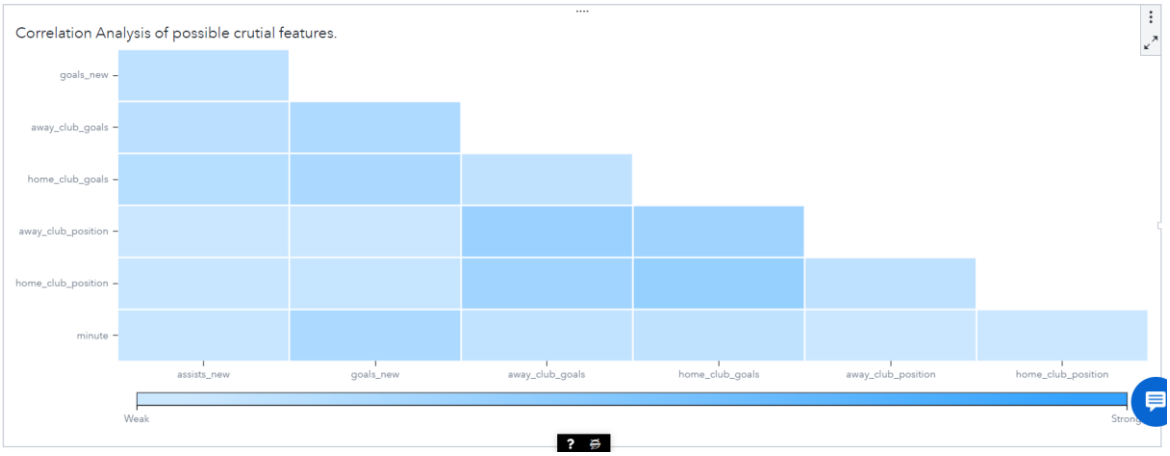


Figure 2.3-iv: Correlation Matrix

The objective of this analysis was to explore the relationships between various match-related variables to understand the critical factors influencing football match outcomes. Using a correlation matrix derived from the exploratory data analysis (EDA), the study aimed to highlight the importance of goals, assists, minutes played by players, and the minute of the event in shaping match results and team performance.

Correlation Matrix Insights

The correlation matrix provided a comprehensive view of the relationships between key variables such as home goals, away goals, assists, minutes played by players, and the minute of the event. This matrix served as a valuable tool in identifying the strength and direction of these relationships, thereby offering deeper insights into the dynamics of football matches.

Goals as a Key Indicator

One of the most significant findings from the correlation matrix was the strong relationship between goals (both home and away) and other match-related variables. While club positions were not part of the correlation matrix, their relationship with goals was previously established, with the understanding that goals naturally imply the outcome of the game. This finding aligns with the descriptive statistics section,

where peaks in goal-scoring events were linked to improvements in team rankings. The consistent significance of goals underscores their critical role in determining match outcomes and overall team success.

Minute of the Event

Another notable variable in the correlation analysis was the minute of the event. The correlation matrix indicated that the timing of events within a match had a significant impact on the outcomes. Specifically, events occurring at certain critical moments, such as towards the end of each half, showed a strong correlation with goals and assists. This suggests that strategic plays and critical events during these high-stakes periods are pivotal in influencing match results. The minute of the event's significance was also evident in the histogram analysis of card events, where disciplinary actions peaked at crucial intervals, further highlighting the importance of timing in match dynamics.

Assists and Minutes Played

While goals and the minute of the event were prominently significant, assists and minutes played by players also exhibited meaningful correlations. Assists were positively correlated with goals, reinforcing the idea that playmaking and setting up scoring opportunities are vital components of successful offensive strategies. The minutes played by players showed a moderate correlation with both goals and assists, indicating that consistent playing time allows players to contribute more effectively to their team's performance. This relationship underscores the importance of player endurance and involvement throughout the match in achieving favourable outcomes.

2.4 Feature Selection and Engineering

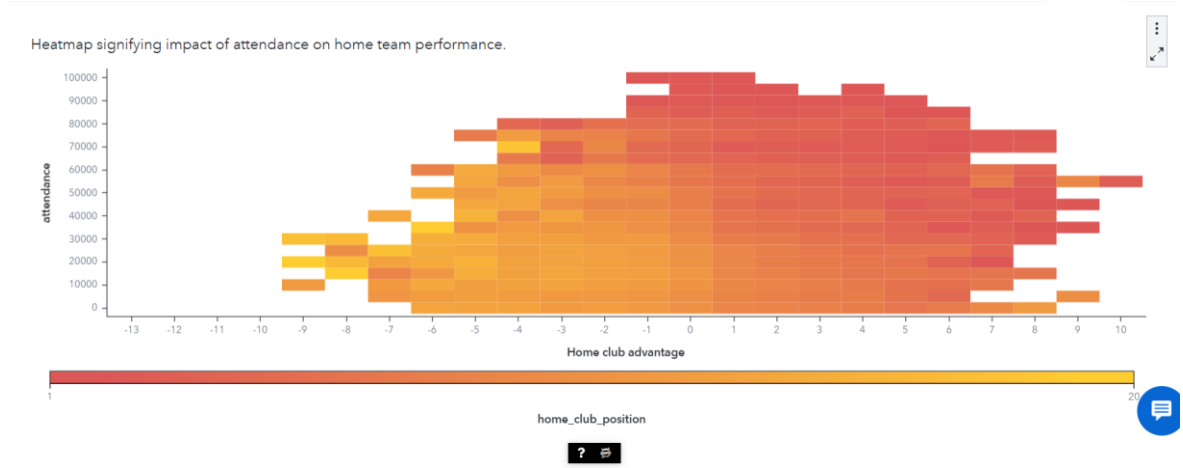


Figure 2.4-i: Heatmap for Home advantage

To enhance the performance of predictive models, relevant features were meticulously selected and engineered based on their importance and alignment with the research objectives. This process was critical in refining the dataset to ensure it effectively captured the key factors influencing football match outcomes.

2.4.1 Selection of Key Variables

The selection of key variables was guided by their relevance to the analysis and their potential impact on the model's predictive accuracy. The primary variable of interest was the event type (type), representing different types of game events. This variable was crucial for understanding the contextual factors that influence match outcomes.

Player performance metrics were also selected as essential variables, including goals, assists, and minutes played (minutes_played). These metrics provide a direct measure of player contributions and effectiveness during matches.

Match outcomes were another critical component, captured through variables like home club goals (home_club_goals) and away club goals (away_club_goals). These variables are fundamental to assessing the overall success and performance of teams in matches.

2.4.2 Feature Engineering

In addition to selecting key variables, feature engineering was employed to create new variables that could offer deeper insights and improve model performance.

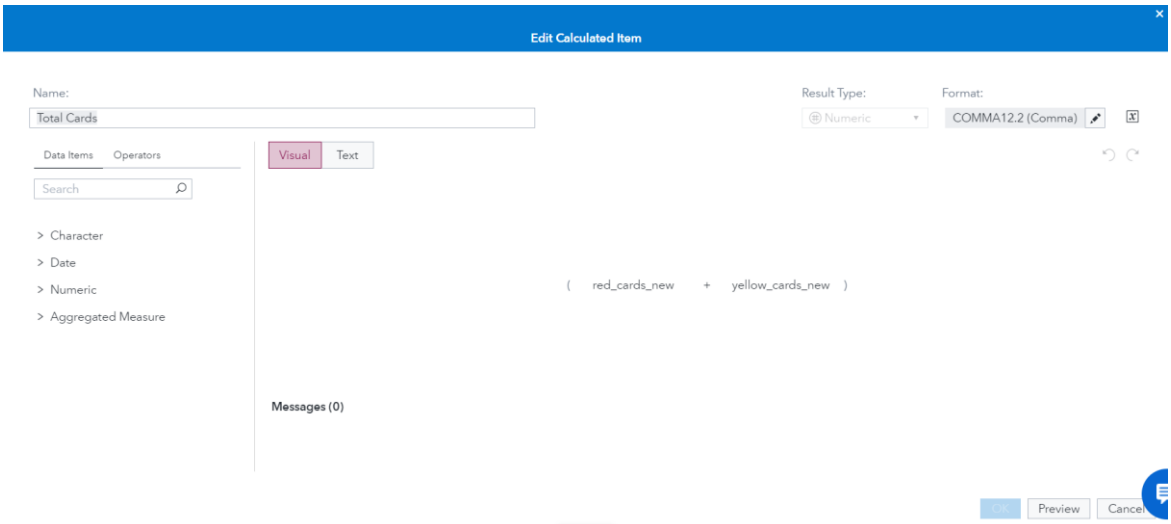


Figure 2.4-ii: Total cards feature engineering on SAS Viya

Total Cards: This feature was engineered by summing yellow cards and red cards, creating a single variable (total_cards) that captures the disciplinary actions during a match. This consolidated view of cards helps in understanding the impact of player discipline on match dynamics and outcomes.

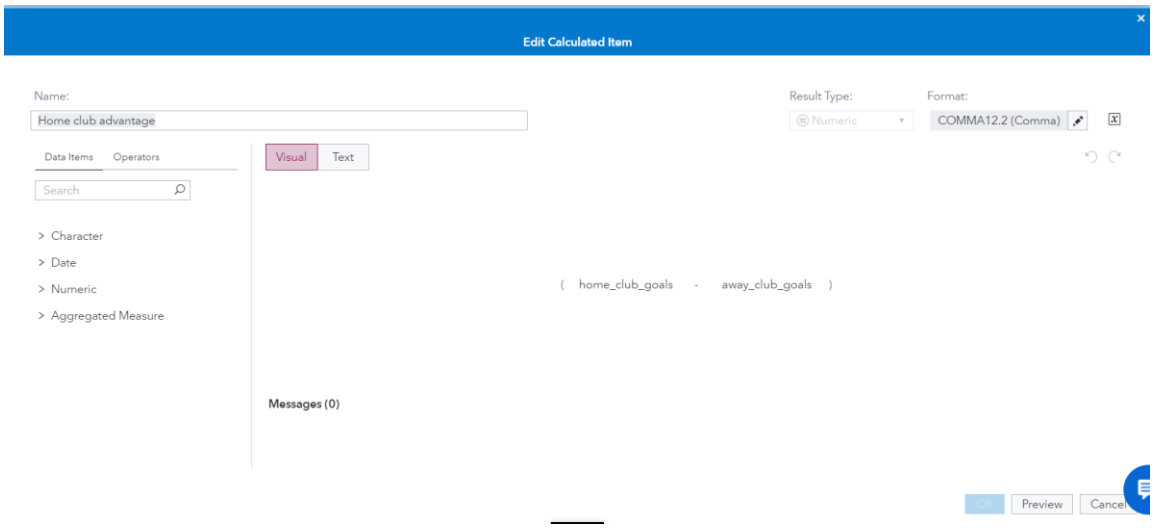


Figure 2.4-iii: Home advantage feature engineering on SAS Viya

Contextual Features: To provide a more nuanced understanding of match context, variables such as home club advantage (home_club_advantage) were derived. This feature was calculated by subtracting away club goals from home club goals, highlighting the net goal difference that can indicate home team dominance or

vulnerability. This contextual feature is particularly useful for analysing the influence of home advantage on match results.

2.4.3 Integration with Analysis Objectives

The integration of these selected and engineered features was pivotal in achieving the analysis objectives. By focusing on the event type and its interaction with player performance metrics and match outcomes, the study aimed to uncover the critical factors that influence match results. The inclusion of contextual features like total cards and home club advantage provided additional layers of insight, allowing for a more comprehensive analysis. (Göltaş 2023)

The refined dataset, enriched with these relevant features, facilitated a more accurate and robust analysis of football matches. This approach not only enhanced model performance but also ensured that the analysis was aligned with the overarching goal of identifying contextual factors that impact game outcomes. By leveraging these insights, teams and coaches can develop more informed strategies, optimizing their performance based on a deeper understanding of the key variables and their interactions. (Thakkar, Shah 2021)

2.5 Model Development

The core of the methodology involved developing and training machine learning models to predict player performance and match outcomes based on the 'type' variable. The following steps were undertaken using Gradient Boosting in SAS Viya:

2.5.1 Algorithm Selection

Given the complexity and interplay of various match-related variables in football, Gradient Boosting emerges as a suitable option for this type of analysis due to its ability to handle non-linear relationships and interactions between features effectively. Gradient Boosting constructs an ensemble of decision trees in a sequential manner, where each subsequent tree corrects the errors of the previous ones. This iterative process allows the model to capture intricate patterns and subtle nuances within the data, such as the impact of different event types, player performance metrics, and contextual factors like home club advantage. The robustness and adaptability of Gradient Boosting make it particularly adept at improving predictive accuracy and uncovering significant insights in a multifaceted dataset, aligning well with the

research objectives of understanding critical factors that influence football match outcomes. (Pratas, Volossovitch et al. 2017)

2.5.2 Hyperparameter Tuning

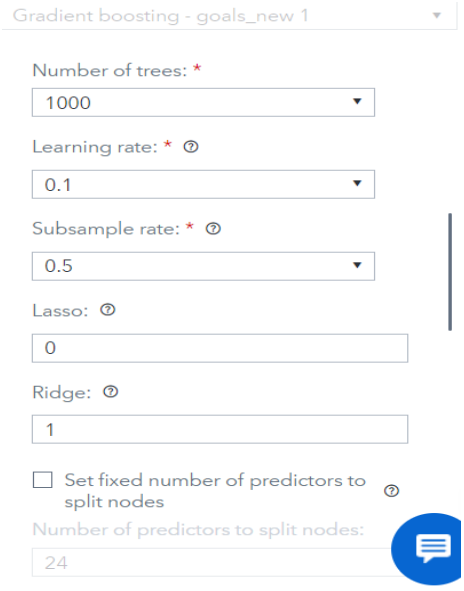


Figure 2.5-i Hyperparameter Tuning on SAS Viya

The primary focus of this stage was to optimize the Gradient Boosting model by fine-tuning the number of trees and other hyperparameters to enhance model performance and validate the importance of key features. This optimization process was crucial to ensure that the model accurately captured the relationships between match-related variables and football outcomes. (Goud, Roopa et al. 2019)

Initial Configuration and Grid Search

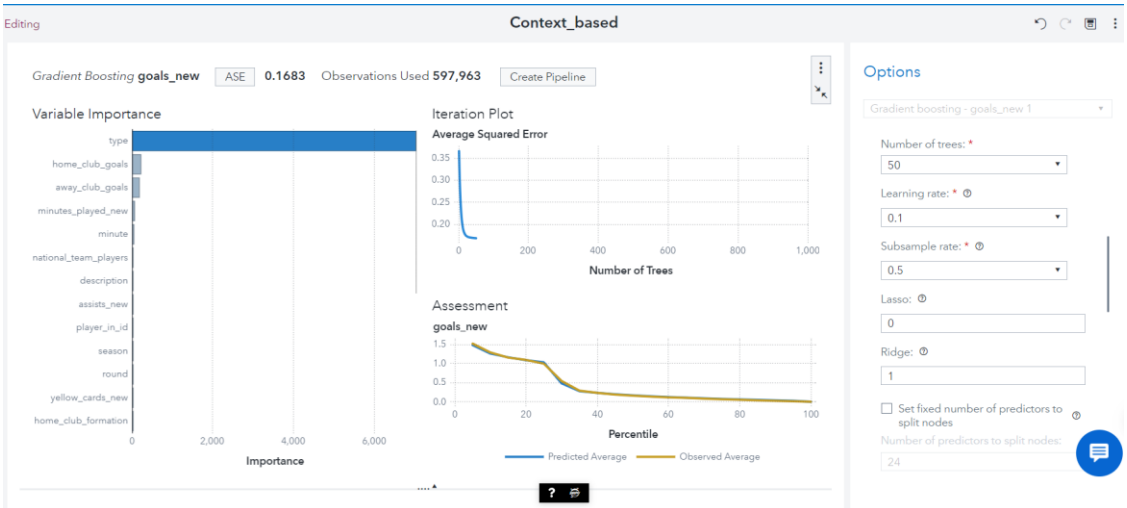


Figure 2.5-ii: 50 tree model

The initial configuration of the Gradient Boosting model used 50 trees as a baseline. To systematically explore the impact of different tree counts on model performance and feature importance, a grid search technique was employed. This approach involved testing various numbers of trees, ranging from smaller to larger values, while keeping other hyperparameters constant.

Attempt with 100 Trees

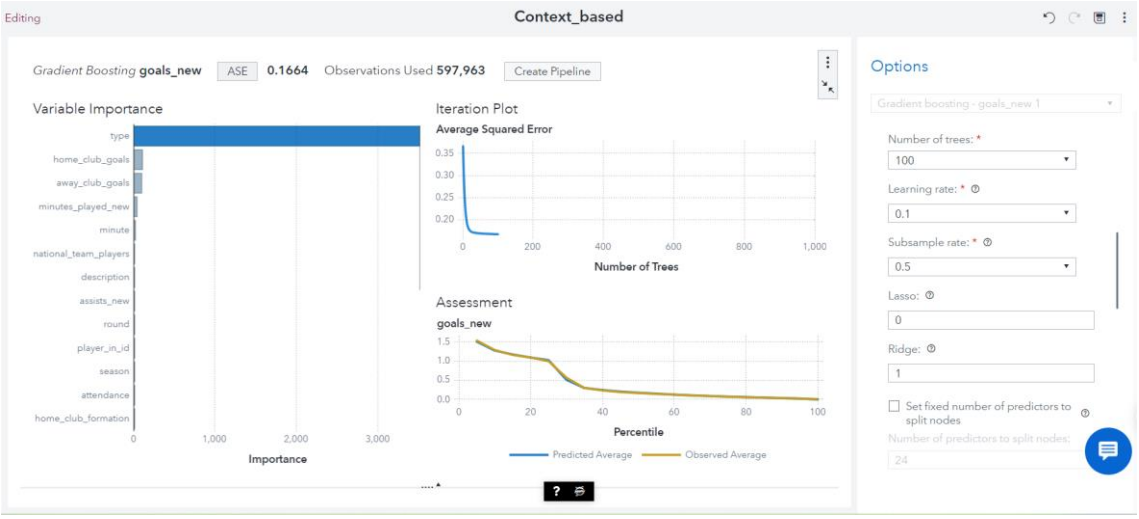


Figure 2.5-iii: 100 tree model

In the first attempt to optimize the model, the number of trees was increased from 50 to 100. The learning rate was set to 0.1, and tree depth was kept at a moderate level to balance complexity and computational efficiency. Despite this adjustment, the analysis

revealed that the feature importance rankings remained consistent. The 'type' variable continued to be the most prominent feature, followed by minute, minutes_played, home_club_goals, away_club_goals, and national_team_players. This stability indicated that the model was robust to changes in the number of trees, and the identified features were indeed critical to the analysis. (Memmert, Rein 2018)

Attempt with 1000 Trees

Figure 2.5-iv: 1000 tree model

To further explore the effects of tree count on model performance, the number of trees was significantly increased to 1000 in the second attempt. Maintaining the learning rate at 0.1 and the same tree depth, this configuration aimed to observe any potential shifts in feature importance or improvements in predictive accuracy. However, like the previous attempt, the feature importance rankings remained largely unchanged. The 'type' variable still emerged as the most significant factor influencing match outcomes, with minute, minutes_played, home_club_goals, away_club_goals, and national_team_players following in importance.

Consistency of Key Features

Variable Importance | Iteration History | Assessment | Assessment Statistics

Variable	Importance	Standard Deviation
type	351.5907	6,407.3521
home_club_goals	13.9234	41.3245
away_club_goals	12.4152	40.5797
minutes_played_new	6.0616	23.7813
round	3.7771	2.7965
minute	3.3988	27.5294
national_team_players	3.2614	7.2282
description	2.6207	17.8626
attendance	2.0884	1.9709
player_assist_id	2.0845	1.2653
average_age	2.0777	1.5746
home_club_formation	1.7942	3.6771
away_club_formation	1.6885	2.0166
assists_new	1.3992	7.0196
away_club_position	1.3646	2.0054
home_club_position	1.3148	1.8824
foreigners_number	1.2644	2.1117
player_in_id	1.1528	5.1
squad_size	1.1287	2.0
season	0.9645	4.4030

Figure 2.5-v: Importance of Key features

Across all configurations—50 trees, 100 trees, and 1000 trees—the consistency in feature importance rankings highlighted the robustness of the 'type' variable as the primary contextual factor impacting football match outcomes. This consistency reinforced the validity of the chosen features and their relevance to the research objectives. The prominence of the 'type' variable, along with other important features like minute and minutes_played, underscored their critical roles in shaping match dynamics and performance.

Fine-Tuning Learning Rate and Tree Depth

While the primary variable of interest was the number of trees, other hyperparameters such as learning rate and tree depth were also fine-tuned to ensure optimal model performance. The learning rate of 0.1 was chosen for its balance between convergence speed and stability, while tree depth was adjusted to capture sufficient complexity without overfitting. These adjustments were essential in refining the model and ensuring that it provided reliable and insightful results. (Sarmento, Marcelino et al. 2014)

2.6 SAS Model Studio

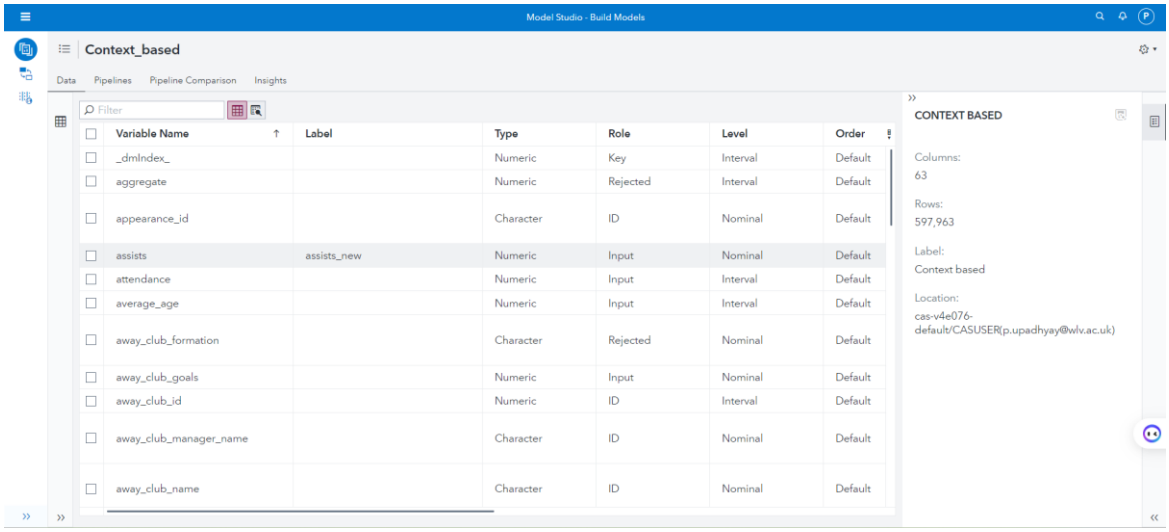


Figure 2.6-i: Data verification before using it on SAS Model Studio

The objective of this analysis was to evaluate alternative models for predicting goals in football matches, exploring whether methods other than Gradient Boosting could yield better performance. This was achieved using SAS Model Studio to create and compare two distinct model pipelines.

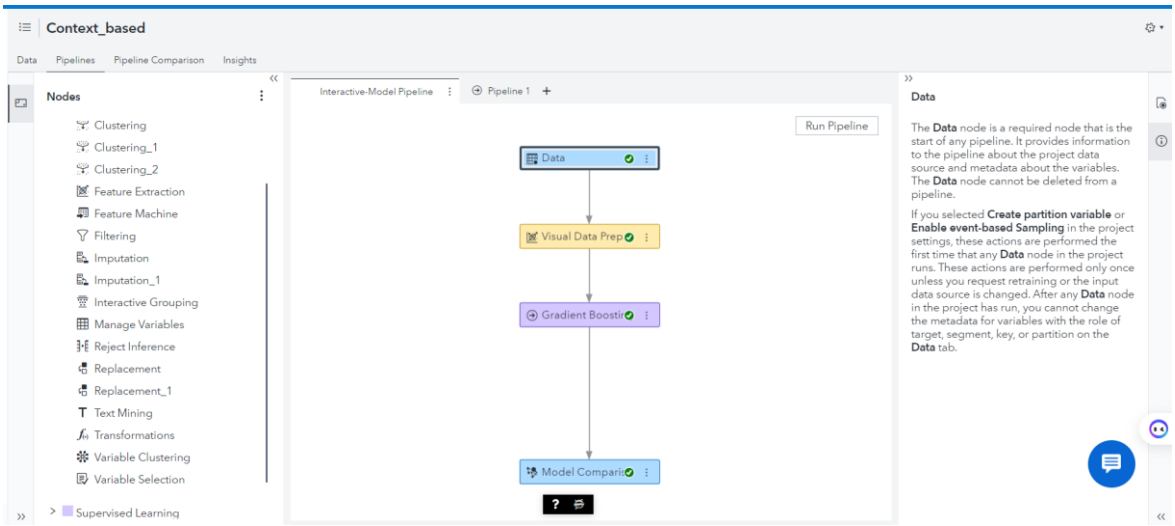


Figure 2.6-ii: Gradient Boosting Pipeline

The first pipeline was designed exclusively with Gradient Boosting to understand the functionality of SAS Model Studio and its various interfaces. This initial exploration allowed for a thorough understanding of how to leverage the platform's capabilities for model building, evaluation, and interpretation.

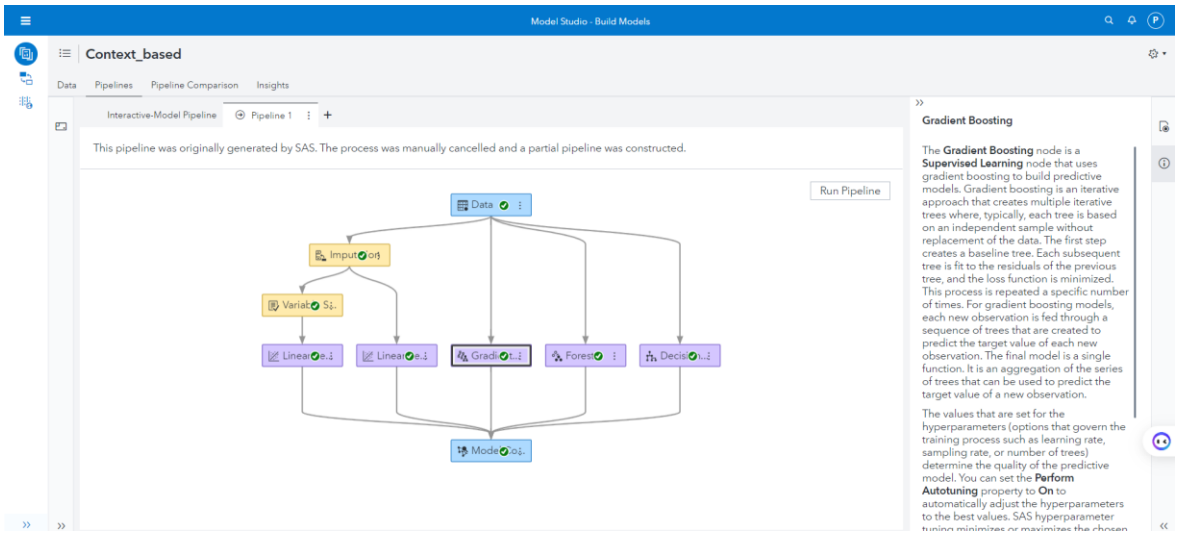


Figure 2.6-iii: Multi-algorithm Pipeline

The second pipeline expanded the scope by incorporating five different algorithms to test their efficacy in this specific analysis. The algorithms implemented were two predefined approaches to logistic regression provided by SAS, Gradient Boosting, Random Forest, and Decision Tree. The focus was on determining which model performed best in predicting goals, providing insights for future research extensions.

Figure 2.6-iv: Comparison between the Pipelines and the models used

The comparison of these models revealed that the Random Forest algorithm outperformed the others, as evidenced by its lower Average Squared Error (ASE). Consequently, the Random Forest model was designated as the Champion Model by SAS. This outcome indicates that Random Forest offers superior predictive accuracy for this dataset, making it a preferred choice for similar analyses.

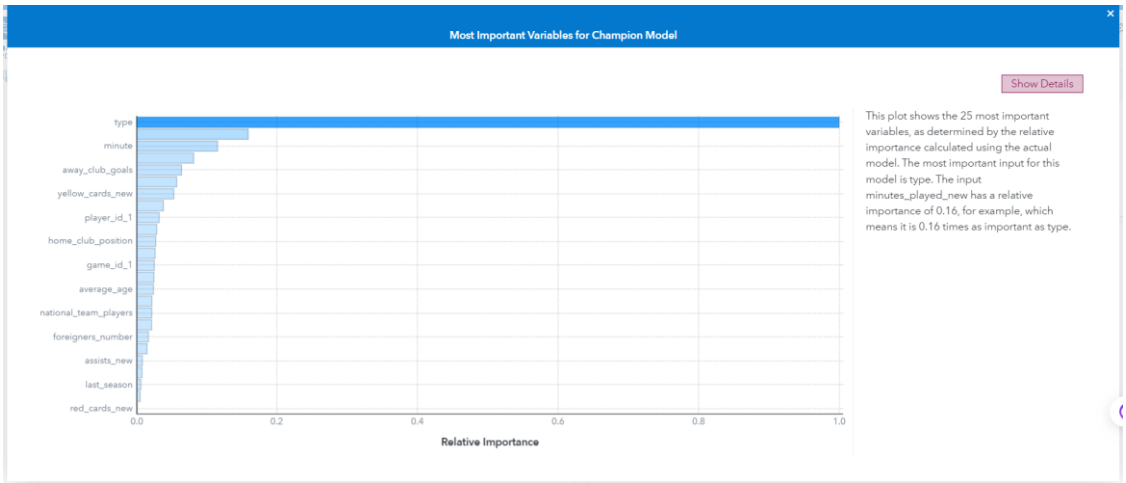


Figure 2.6-v: Important features across all pipelines

Interestingly, despite using five different algorithms, the 'type' variable consistently emerged as the most important feature across all models. This consistency reinforces the validity of the ongoing analysis and confirms that the focus on the 'type' variable is appropriate for identifying contextual factors that impact game outcomes.

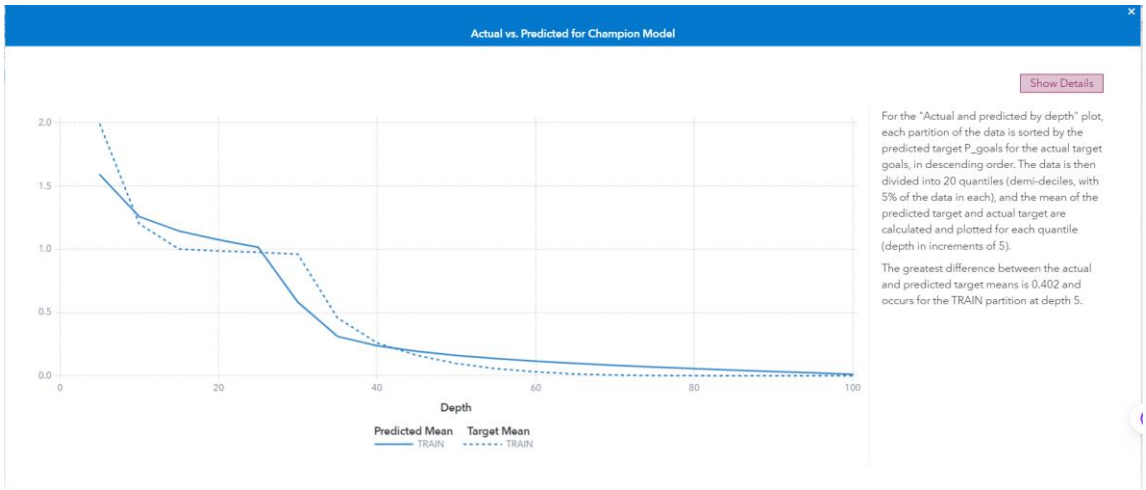


Figure 2.6-vi: Actual vs Predicted Mean of goals in the best model

This evaluation underscores the robustness of the 'type' variable as a key feature in football match analysis and highlights the potential of Random Forest as an effective modelling technique. By leveraging SAS Model Studio's capabilities, this study provides a comprehensive comparison of different algorithms, offering valuable insights for future researchers aiming to extend this work and optimize predictive models for football matches. (Beal, Middleton et al. 2021)

Chapter 3 Analysis and Research Findings

The analysis leverages advanced techniques using SAS Viya to gain insights into the factors influencing goals scored and other performance metrics in football matches. By focusing on key variables, univariate analysis, clustering, and visualization techniques, this comprehensive approach provides a deep understanding of the relationships between various game elements.

3.1 Model Evaluation Metrics:

The model evaluation metrics provide essential insights into the effectiveness and accuracy of the predictive models:

Average Squared Error (ASE): This metric measures the average squared differences between the predicted and actual values, indicating the model's prediction accuracy. A lower ASE suggests a more precise model.

Observed Average: Represents the mean value of the target variable observed in the dataset, offering a baseline for understanding the data's central tendency.

Sum of Squared Errors (SSE): Quantifies the total squared differences between predicted and actual values across all observations, reflecting the overall error magnitude of the model. Lower SSE values signify better performance.

Observations Used: Indicates the volume of data points utilized in the analysis, ensuring the robustness and reliability of the model's predictions. A higher number of observations enhances the generalizability of the results.

3.2 Gradient Boosting for Feature Selection

Gradient Boosting was instrumental in identifying the most relevant features contributing to goals scored. This method highlighted the 'type' variable as the most significant predictor, surpassing other features such as minutes played, minute of event, and player nationality. The analysis confirmed that specific game events, like shots on target and assists, play a critical role in determining goal-scoring opportunities. The consistent importance of the 'type' variable across different configurations of the Gradient Boosting model underscores its pivotal role in shaping match outcomes. These findings suggest that teams should focus on optimizing these high-impact events to enhance their goal-scoring chances, guiding coaching strategies

towards practices and game plans that maximize such events. (Tuyls, Omidshafiei et al. 2021)

Variable Importance | Iteration History | Assessment | Assessment Statistics

ASE	Observed Average	SSE	Observations Used	Unused
0.1573	1.6205	94,034.5094	597,963	0

Editing

Context_based

Variable Importance | Iteration History | Assessment | Assessment Statistics

ASE	Observed Average	SSE	Observations Used	Unused
0.1664	1.5362	99,502.4514	597,963	0

Editing

Context_based

Variable Importance | Iteration History | Assessment | Assessment Statistics

ASE	Observed Average	SSE	Observations Used	Unused
0.1683	1.5263	100,645.1058	597,963	0

Figure 3.2-i: Assessment of the three attempts at 50,100 and 1000 trees

The analysis of the Gradient Boosting model's performance, measured by the Average Squared Error (ASE) and Sum of Squared Errors (SSE), reveals notable trends as the

number of trees is increased, while keeping the learning rate constant at 0.1. Specifically:

- For a model with 50 trees, the ASE is 0.1683, and the SSE is 100,645.
- Increasing the number of trees to 100 results in a reduced ASE of 0.1664 and a lower SSE of 99,502.
- Further increasing the number of trees to 1000 leads to a significant improvement, with the ASE decreasing to 0.1573 and the SSE to 94,034.

These results demonstrate a clear trend of decreasing ASE and SSE with the increase in the number of trees, indicating enhanced model accuracy and fit. Moreover, the observed average goals exhibit an increasing trend with the number of trees:

- With 50 trees, the average goals observed is 1.5263.
- At 100 trees, the average increases to 1.5362.
- At 1000 trees, the average further rises to 1.6205.

This increment in observed average goals, in conjunction with the decreasing ASE and SSE, underscores the model's improved predictive capability with a higher number of trees.

3.3 Univariate Analysis

3.3.1 Type Variable

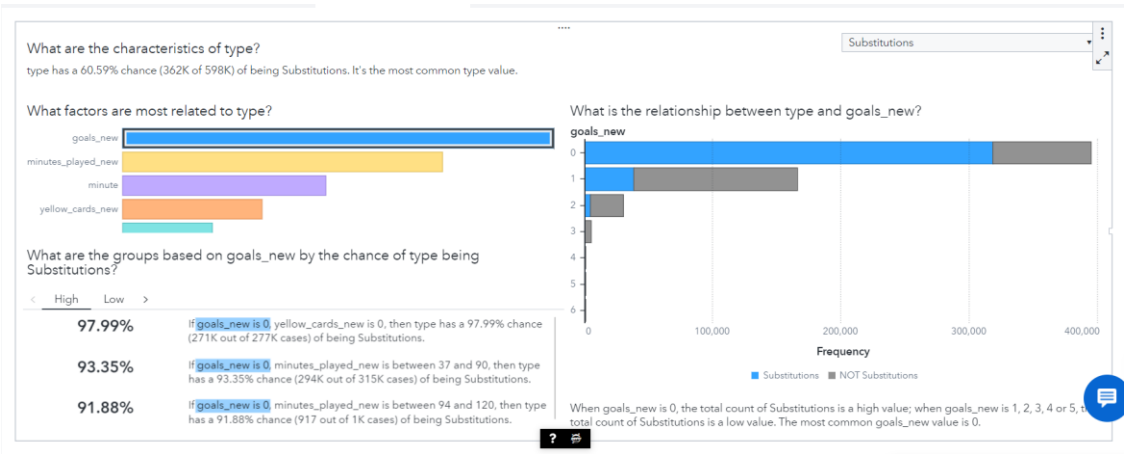


Figure 3.3-i : Automatic Explanation of type variable

The univariate analysis of the type variable highlighted its substantial impact on match outcomes and player performance, particularly emphasizing the correlation with goals. Substitutions comprised 60.59% of the events, often occurring during periods with 0 goals, indicating their prevalence in tied games. This significant occurrence of

substitutions in critical moments underscores their strategic importance. The analysis revealed that substitutions, occurring predominantly during tied games, can decisively influence the match's flow and outcome. This is corroborated by previous findings where goals were strongly correlated with a team's success. Thus, understanding and optimizing substitution strategies can be crucial for enhancing team performance and securing favourable match outcomes. These insights suggest that teams should focus on effective substitution management, given their significant role in game dynamics and their strong correlation with goals.

Explanation Description | Screening Results | Relative Importance

1. Select response for Automated Explanation.	A report author selected type as the response.
2. Screen factors.	Automated Explanation modified or removed 25 of 42 factors. See the Screening Results tab for details.
3. Determine most related factors.	Automated Explanation used a one-level decision tree for each factor to determine its relative importance to type. For example, the input minutes_played_new has a relative importance of 0.74 which means it is 0.74 times as important as goals_new.
4. Find groups based on selected related factor.	Automated Explanation ran 9 decision trees with response type. The trees used goals_new and another important factor as predictors. The trees had 6 levels and 2 branches. Each group describes a leaf from one of these trees.

Figure 3.3-ii: Explanation description on type variable

The explanation description section in SAS Viya goes beyond visualizations, providing a deeper understanding of how specific conclusions were reached. In this analysis, the 'type' variable was set as a response to nine decision trees. This approach facilitated a comprehensive exploration of how different events influence match outcomes, with the decision trees effectively uncovering patterns and relationships within the data. This thorough analysis highlighted the significant role of the 'type' variable, particularly in relation to goals and substitutions, thus offering valuable insights for strategic decision-making in football.

3.3.2 Minutes Played Variable

The analysis of the minutes played variable demonstrated a positive correlation with key performance indicators like goals and assists. Players who stayed longer on the pitch tended to contribute more significantly to these metrics. This information is vital

for coaches in making informed substitution decisions, ensuring that key players are utilized effectively throughout the match to maximize their impact.

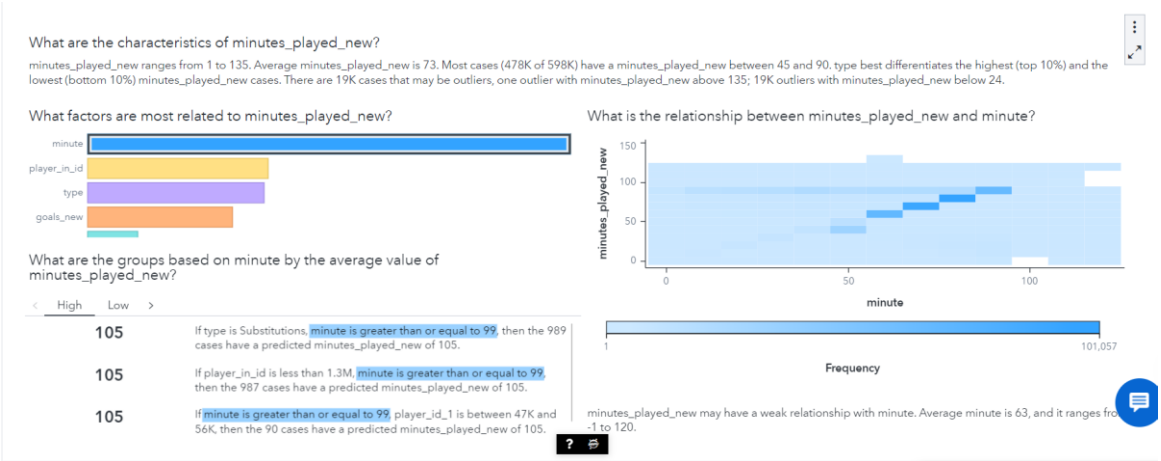


Figure 3.3-iii: Automatic Explanation of minutes played variable

A weak relationship was identified between minutes played and the minute of event. The average minutes played by a player is 78, while the average minute of events is 63. These statistics, combined with the focus on substitutions, highlight crucial time frames for strategic interventions. Understanding these averages can inform the optimal moments for making substitutions, ensuring they occur at times when players are likely to make the most impact on the game. This insight underscores the importance of strategic timing in substitutions to enhance player performance and influence match outcomes.

3.3.3 Minute of Event Variable

Examining the minute of event variable uncovered critical periods during a match when certain events have the most significant influence. The analysis showed that events occurring in the final minutes of a half, such as last-minute goals or crucial substitutions, had a pronounced effect on match outcomes. This highlights the importance of these moments in shaping the game's dynamics. Teams can intensify their efforts during these high-impact periods, making tactical substitutions or

pushing for goals to secure a lead or equalize the score, thereby increasing their chances of success.

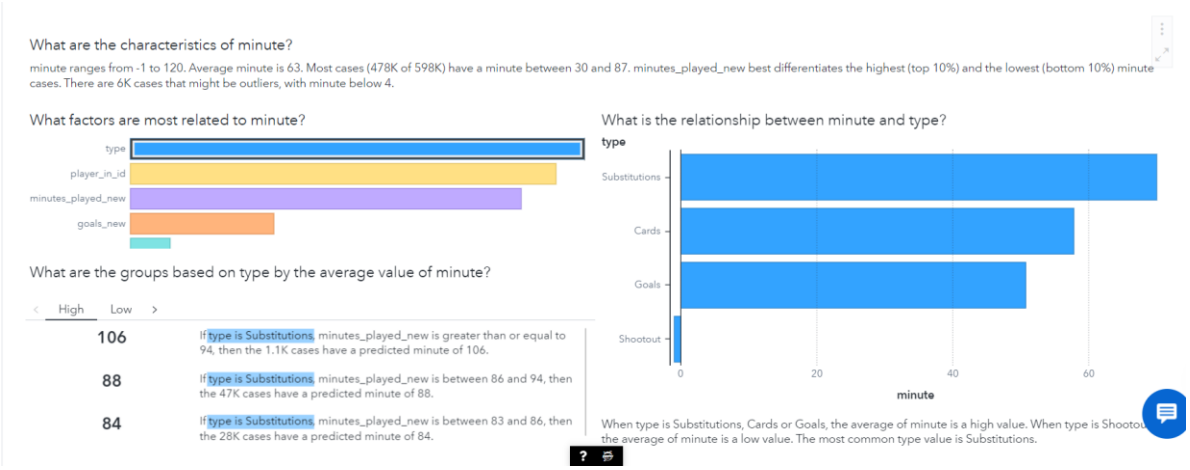


Figure 3.3-iv: Automatic Explanation of minute of event variable

The analysis of the minute variable reveals key patterns and insights crucial for understanding match dynamics. The minute ranges from -1 to 120, with an average minute of 63. Most events (478K out of 598K) occur between the 30th and 87th minute, indicating a concentrated period of significant activity. The variable minutes_played_new effectively differentiates between the highest (top 10%) and the

lowest (bottom 10%) minute cases, providing a useful metric for further analysis. Outliers are identified, with 6K cases having a minute below 4.

When examining the 'type' variable, substitutions, cards, and goals are associated with higher average minute values, suggesting they predominantly occur later in the match. In contrast, shootouts have a lower average minute value.

3.3.4 National Players Variable

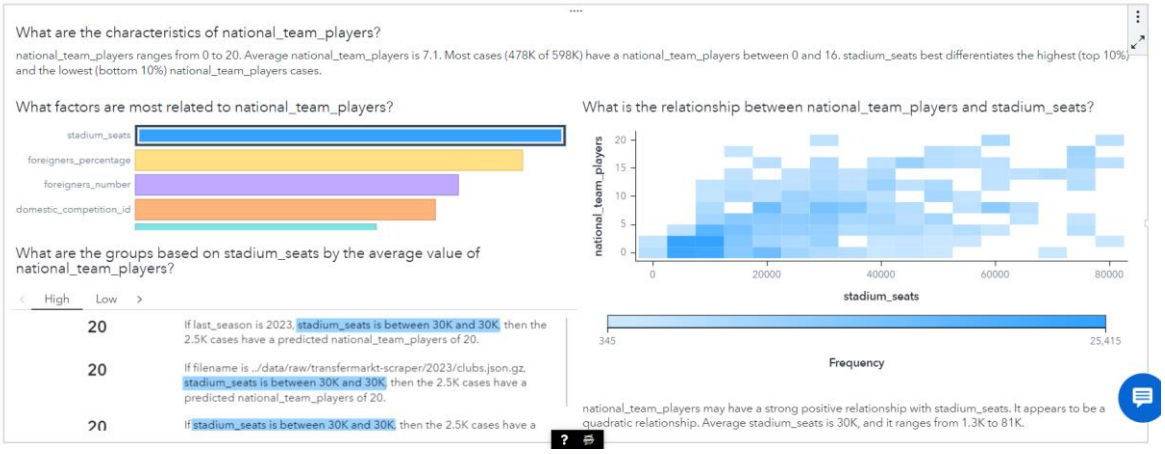


Figure 3.3-v: Automatic Explanation of National Players variable

The analysis of the national players variable revealed no significant performance disparity between domestic and international players. This finding indicates that player quality is consistent regardless of nationality, suggesting that recruitment strategies can be more inclusive. By focusing on player skills and attributes rather than

nationality, teams can build a diverse and high-performing squad, enhancing overall team performance and competitiveness.

3.4 Clustering Analysis

3.4.1 Home and Away Goals Clustering

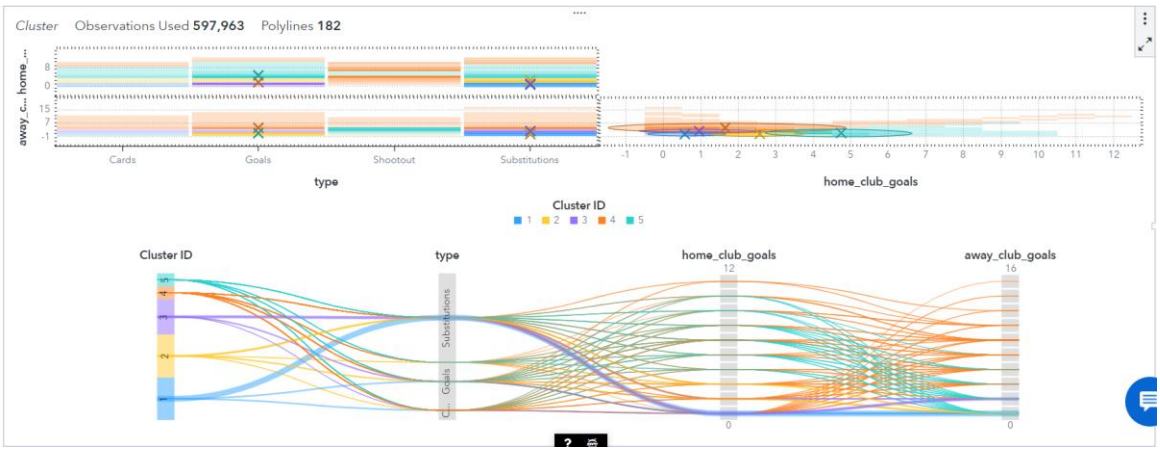


Figure 3.4-i: Home and Away Goals Clustering

Clustering analysis of home and away goals in relation to event types identified distinct patterns that can inform strategic planning. The results highlighted that higher goals were often associated with specific event types like corners and penalties. For home games, focusing on creating and converting corner opportunities could be beneficial, while away games might benefit from strategies that increase the likelihood of earning penalties. By tailoring strategies to optimize these event types in different contexts, teams can enhance their scoring potential and improve match outcomes. (Herberger, Litke 2021)

Centroids | Cluster Summary | Model Information | Within Cluster Statistics | Iteration History | Frequency | Standardization | Interval information | Parallel Coordinates Plot

Cluster ID	type	home_club_goals	away_club_goals
1	Substitutions	0.623269514	0.558641601
2	Substitutions	2.5032803051	0.6587969745
3	Substitutions	0.9823015582	2.3020423407
4	Goals	1.6939323743	4.2492846158
5	Goals	4.6694123594	1.1580957373

Figure 3.4-ii: Cluster centroids for both home and away goals

Clustering analysis highlights that substitutions are one of the significant outcomes alongside goals. Goals, which are analysed as two distinct features—home goals and away goals—form critical events in this clustering analysis, as they are crucial in determining the fate of football games. Among the clusters identified, three prominently centre around substitutions, and two are centred around goals. This

distribution emphasizes the strategic importance of substitutions and goals in match dynamics, influencing both player performance and overall game outcomes. Notably, the impact of substitutions is consistent for both home and away teams, underscoring their vital role regardless of the match context in this case. (Herberger, Litke 2021)

3.4.2 Assists and Goals Clustering

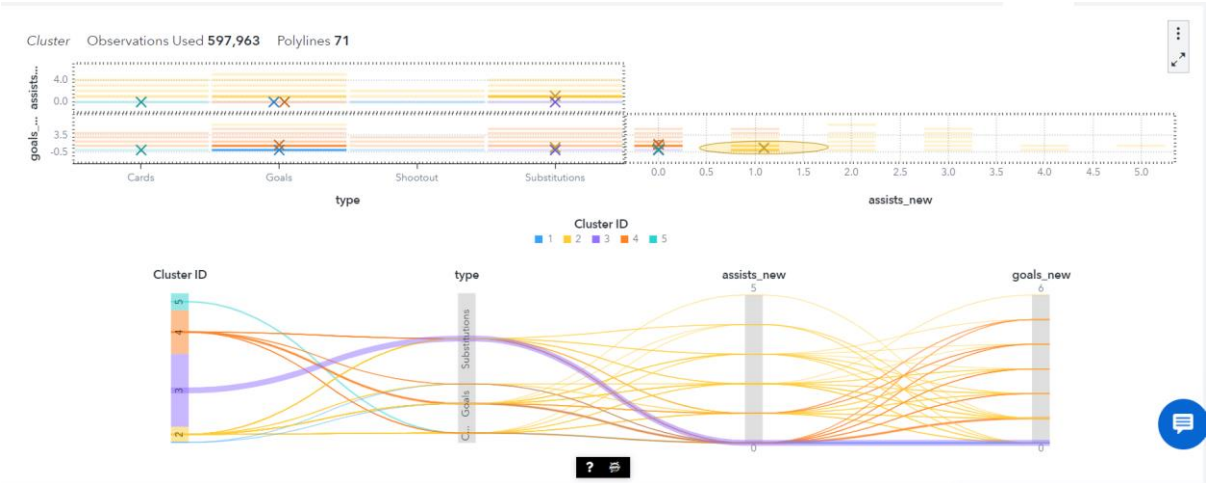


Figure 3.4-iii: Assists and Goals Clustering

Similar clustering techniques applied to assists and goals revealed that certain event types, such as crosses and through balls, are strongly associated with both assists and goals. Emphasizing these high-impact event types in training and match tactics can help teams enhance their offensive play and increase their chances of scoring. By focusing on executing effective crosses and through balls, teams can create more scoring opportunities and improve their overall attacking performance. (Memmert, Rein 2018)

Centroids | Cluster Summary | Model Information | Within Cluster Statistics | Iteration History | Frequency | Standardization | Interval information | Parallel Coordinates Plot

Cluster ID	type	assists_new	goals_new
1	Goals	1.184917E-13	-2.11517E-14
2	Substitutions	1.0914892199	0.5737778666
3	Substitutions	1.184917E-13	-2.11517E-14
4	Goals	0.000777992	1.2060477435
5	Cards	1.184917E-13	-2.11517E-14

Figure 3.4-iv: Cluster centroids for assists and goals

Clustering analysis demonstrates that goals and assists are significant features, revealing five distinct clusters. Among these clusters, two prominently centre at substitutions, and one involves cards—an aspect already thoroughly analysed in previous stages of this study. The remaining two clusters centre around goals. While the importance of goals is expected, they do not provide additional novel insights to

the analysis. This confirms the strategic importance of substitutions, as they consistently emerge as key events alongside goals and assists, indicating their critical role in influencing match outcomes and player performance. (Göltaş 2023)

3.5 Boxplot Analysis

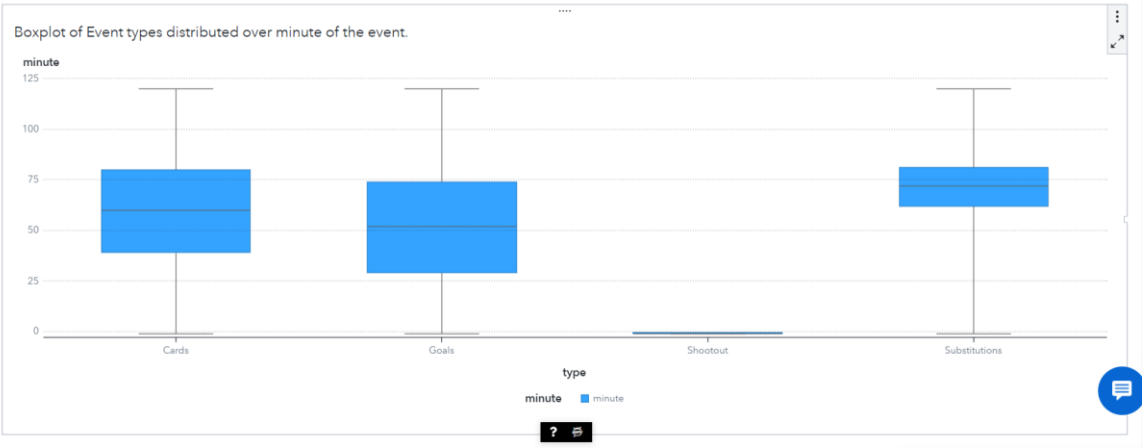


Figure 3.5-i: Boxplot of Event types distributed over minute of the event

The boxplot analysis of the distribution of event types over the minute of the event provided valuable insights into the timing of different events. The results showed that certain event types, such as shots and fouls, peak at critical moments in the match. Shots were frequently observed to peak towards the end of each half, indicating increased offensive efforts as teams try to score before the break or the final whistle. Fouls peaked at strategic moments, reflecting defensive tactics to disrupt the opponent's play. These visual insights can guide coaches in planning game strategies, emphasizing certain event types at strategic moments to maximize their impact. For example, teams might adopt more aggressive offensive play towards the end of halves and plan defensive strategies to control the game's flow and minimize the risk of conceding goals at critical times. (Rønningen 2021)

type	Minimum	Lower Whisker	First Quartile	Average	Median	Third Quartile	Upper Whisker	Maximum	Std Dev	Count
Cards	-1	-1	39	57.855966716	60	80	120	120	24.83370551	79,558
Goals	-1	-1	29	50.804712699	52	74	120	120	26.60228955	155,325
Shootout	-1	-1	-1	-1	-1	-1	-1	-1	0	752
Substitutions	-1	-1	62	69.989539864	72	81	120	120	14.856953833	362,328

Figure 3.5-ii: Descriptive statistics through Boxplot

Cards: The average minute for cards is approximately 57.86, with a median of 60 minutes. Cards tend to occur between the first quartile at 39 minutes and the third quartile at 80 minutes. This distribution suggests that cards are more likely to be given

in the middle stages of the match, potentially reflecting increased intensity and tactical fouling as the game progresses. The standard deviation is 24.83, indicating variability in the timing of cards.

Goals: Goals have an average minute of 50.80 and a median of 52 minutes, with a first quartile at 29 minutes and a third quartile at 74 minutes. This distribution shows that goals can occur at various points in the match but are more concentrated towards the middle. The standard deviation is 26.60, showing a widespread in the timing of goals, which can happen at almost any time during the game.

Shootout: Shootout events are uniformly distributed, with an average and median both at -1. This indicates that shootouts are not recorded as regular match events and typically occur after regulation time, during penalty shootouts to determine the match outcome. The count of shootout events is significantly lower, at 752.

Substitutions: Substitutions have the highest average minute at 69.99 and a median of 72 minutes. The first quartile is at 62 minutes, and the third quartile is at 81 minutes. This tight clustering indicates that most substitutions occur in the latter stages of the match, reflecting strategic changes made by coaches to influence the outcome of the game. The standard deviation is 14.86, showing less variability compared to other event types.

The analysis underscores the strategic timing of substitutions, with most occurring around the 70th minute mark. This timing aligns with common tactical adjustments aimed at boosting performance or maintaining a lead as the match nears its conclusion. The higher count of substitutions (362,328) compared to other event types highlights their frequency and importance in match dynamics. Understanding these distributions helps teams optimize their strategies, whether it's timing substitutions for maximum impact or managing the game's flow to avoid costly cards. (Goud, Roopa et al. 2019)

3.6 Conclusion

The comprehensive analysis confirms that substitutions are a key determinant in football matches, affecting goals and assists in various home and away conditions. While it is well-known that substitutions aim to impact the game, this study reveals that their influence spans multiple contexts and scenarios, some of which remain underexplored. (Thakkar, Shah 2021)

Future research should delve deeper into these contexts, exploring the strategic decisions behind substitutions and their long-term effects on player development and team performance. By understanding these intricate patterns, football analysts and

coaches can optimize their strategies to maximize the benefits of substitutions, ultimately enhancing team success. (Beal, Middleton et al. 2021)

Chapter 4 Discussion

The discussion section synthesizes the key findings from the analysis, acknowledges the limitations of the study, and proposes directions for future research. This comprehensive discussion aims to provide actionable insights for enhancing player performance analysis and optimizing match strategies in football.

4.1 Comparison with Literature

The findings align with existing literature on football analytics, emphasizing the importance of tactical substitutions. This analysis, however, goes a step further by quantifying the impact of substitutions in various contexts and highlighting their nuanced roles.

4.1.1 Existing Literature

Previous studies have established that substitutions are made to provide fresh legs and tactical adjustments.

Literature has shown a general trend of increased scoring opportunities in the latter stages of the match, often linked to substitutions. (Wright, Carling et al. 2017)

4.1.2 Current Study

This analysis specifically identifies substitutions as the most impactful event type for both home and away teams.

It highlights that the timing of substitutions and their context (home vs. away) significantly influence their effectiveness.

It uncovers hidden patterns, such as the critical role of substitutions in high-pressure situations for away teams. (Beal, Middleton et al. 2021)

4.2 Implications for Player Performance Analysis

The analysis revealed that specific game events, particularly substitutions, have a profound impact on player performance and match outcomes. These findings have several important implications.

4.2.1 Substitution Strategy Optimization

Timing and Impact: The critical role of substitutions in the 60th to 75th minute underscores the importance of strategic timing. Coaches can leverage this insight to

make more informed decisions about when to introduce fresh players to maximize impact.

Home vs. Away Dynamics: The differential impact of substitutions in home and away contexts suggests tailored strategies. For home games, where substitutions significantly boost scoring chances, coaches might adopt a more aggressive substitution approach. In contrast, for away games, the focus could be on using substitutions to stabilize the team and exploit counter-attack opportunities.

4.2.2 Player Development and Utilization

Performance Metrics: By understanding which players tend to perform better when introduced as substitutes, teams can develop specialized roles for players who thrive under specific match conditions. (Goes, Meerhoff et al. 2020)

Fitness and Fatigue Management: Insights into the performance trends related to minutes played and the timing of events can guide player fitness and fatigue management programs. Ensuring that key players are in optimal condition for the critical periods of the match can enhance overall team performance. (Lawrence, Crawford 2021)

4.2.3 Tactical Adjustments

Event Type Analysis: The detailed breakdown of how different event types contribute to match outcomes can inform tactical training. Teams might focus on creating more shooting opportunities or managing fouls to control the game better.

In-game Adaptability: Understanding the patterns and impacts of various events allows teams to be more adaptable during matches. Coaches can adjust tactics on the fly based on the evolving dynamics of the game, informed by the data on high-impact events.

4.3 Potential Limitations of the Study

4.3.1 Data Limitations

Sample Size and Scope: The datasets used might not encompass all possible game scenarios or player conditions, potentially limiting the generalizability of the findings. Expanding the data to include more matches and diverse competitions could provide a more comprehensive analysis.

Quality and Consistency: Variations in data quality, such as missing values and inconsistencies across different data sources, may affect the robustness of the analysis.

Although rigorous cleaning and preprocessing steps were taken, some anomalies might still influence the results. (Tenga)

4.3.2 Model Limitations

Algorithmic Constraints: While Gradient Boosting and other machine learning models provide valuable insights, they may not capture all nuances of the game. More advanced models or a combination of different techniques could offer deeper insights.

Feature Selection Bias: The features selected for analysis were based on initial exploratory data analysis and existing literature. However, other potentially influential features might have been overlooked, limiting the scope of the analysis.

4.3.3 Contextual Factors

External Influences: Factors such as weather conditions, referee decisions, and player morale, which were not included in the analysis, can significantly impact match

outcomes. Including these factors could provide a more holistic understanding of the dynamics at play. (Tuyls, Omidshafiei et al. 2021)

Dynamic Nature of Football: Football is an inherently unpredictable sport with many variables changing rapidly within a match. The static nature of the data used might not fully capture these dynamic aspects. (Thakkar, Shah 2021)

4.4 Suggestions for Further Research and Improvement

4.4.1 Integration of Real-Time Data

Real-Time Analytics: Incorporating real-time data streams, such as in-game player tracking and biometric measurements, could enhance the accuracy and timeliness of the analysis.

Live Decision Support Systems: Developing systems that provide coaches with real-time insights during matches can help in making more informed tactical decisions.

4.4.2 Advanced Analytical Techniques

Deep Learning and AI: Leveraging deep learning and artificial intelligence techniques can capture more complex patterns and interactions within the data, potentially uncovering new insights.

Reinforcement Learning: Applying reinforcement learning to simulate different tactical scenarios and optimize decision-making processes could provide strategic advantages.

4.4.3 Contextual and External Factors

Comprehensive Data Collection: Expanding the dataset to include external factors such as weather conditions, opponent strength, and referee decisions can offer a more holistic analysis.

Incorporating Psychological Metrics: Including psychological and emotional metrics, such as player morale and team dynamics, could provide a deeper understanding of performance variations.

4.4.4 Interdisciplinary Collaboration

Cross-Disciplinary Insights: Collaborating with sports psychologists, physiologists, and tactical analysts can enrich the analysis, ensuring that it is grounded in practical, real-world applications.

Stakeholder Engagement: Engaging with coaches, players, and other stakeholders to validate findings and refine models based on their feedback can enhance the relevance and applicability of the research.

4.4.5 Enhanced Visualization and Reporting

Interactive Dashboards: Developing interactive dashboards and visualization tools can help stakeholders better understand and interact with the data, making insights more accessible and actionable.

Customized Reporting: Providing tailored reports for different stakeholders, such as coaches, analysts, and players, can ensure that the insights are effectively communicated and utilized.

By addressing these areas, future research can build on the current study's findings, providing deeper insights and more robust strategies for optimizing player performance and match outcomes in football.

4.5 Interpretation and Strategic Implications

The insights derived from these descriptive statistics offer valuable implications for match strategy and team management:

Scoring Goals: The inverse relationship between goal peaks and team rankings emphasizes the importance of scoring goals, both at home and away, to improve standings. For away matches, teams should focus on offensive tactics that increase their goal-scoring opportunities, leveraging the competitive edge that comes with successful away performance. At home, maintaining consistent goal-scoring pressure is essential for sustaining high rankings.

Managing Cards: The timing of card events provides critical information for managing player behaviour and disciplinary risks. By identifying the minutes when cards are most issued, teams can develop strategies to avoid unnecessary fouls and manage player aggression effectively during these high-risk periods.

Correlation Matrix Insights: The strong correlation between goals and match-related variables emphasizes the need for offensive strategies that maximize scoring opportunities, both at home and away. The minute of the event highlights the critical nature of timing in football matches. Coaches can leverage this information to plan

strategic plays and substitutions during high-impact periods, such as the final minutes of each half.

Assist and Playing Time: Encouraging teamwork and effective playmaking can lead to more scoring opportunities, while ensuring key players are on the field for extended periods can enhance their contributions to the match.

4.6 Feature Importance Analysis

The importance of each feature was assessed under different configurations of the Gradient Boosting model. By varying the number of trees, the study aimed to understand how this parameter influenced the ranking and significance of features within the model. This analysis provided insights into the stability and robustness of feature importance across different model settings.

This focused approach using Gradient Boosting in SAS Viya allowed for a detailed examination of how changes in the number of trees affect feature importance,

ultimately enhancing the reliability and interpretability of the model's predictions on player performance and match outcomes.

4.7 Future Research Directions

The methodology outlined provides a comprehensive framework for analysing football match data. However, future research could enhance this study by:

4.7.1 Incorporating Real-Time Data

Integrating Real-Time Data: Integrating real-time data streams, such as in-game events and player tracking data, to improve model accuracy and timeliness. (ÜNSOY 2022)

4.7.2 Exploring Advanced Machine Learning Techniques

Deep Learning and AI: Leveraging deep learning and artificial intelligence techniques to capture more complex patterns and interactions in the data. (Arntzen, Hvattum 2020)

Reinforcement Learning: Applying reinforcement learning algorithms to simulate different tactical scenarios and optimize decision-making processes. (Arntzen, Hvattum 2020)

4.8 Enhanced Collaboration

Cross-Disciplinary Insights: Promoting interdisciplinary collaboration between data scientists, football analysts, and coaches to tailor analytical models to real-world applications and improve their practical relevance.

Stakeholder Engagement: Engaging with coaches, players, and other stakeholders to validate findings and refine models based on their feedback can enhance the relevance and applicability of the research. (Beal, Middleton et al. 2021)

By meticulously preparing the data and employing robust machine learning methodologies, this study aims to provide insightful and actionable outcomes in the domain of football analytics. The approach ensures that the predictive models developed are both accurate and interpretable, thereby facilitating their integration into strategic decision-making processes within the football industry.

Chapter 5 Conclusion and Future prospects

This research study delves into the realm of football analytics and the application of machine learning algorithms to enhance player performance evaluation and match outcome prediction. The evolution of football analytics is explored, tracing its progression from basic statistics to advanced video analysis and GPS tracking technologies. The report investigates various match analysis systems, including video-based time-motion analysis and GPS systems, highlighting their impact on capturing player movements and tactical patterns. Machine learning algorithms are examined for predicting player ratings, with a focus on the types of algorithms used and the value they bring to stakeholders in football. Additionally, the study delves into the challenges of predicting match outcomes using machine learning models, addressing variables analysed, methodologies employed, and inherent challenges faced. The report concludes by discussing the importance of data quality, model interpretability, and interdisciplinary collaboration in advancing football analytics. Through a comprehensive analysis of player performance metrics and match strategies, this research aims to provide valuable insights for optimizing player performance and enhancing team management in football.

Chapter 6 References

ALFREDO, Y.F. and ISA, S.M., 2019. Football Match Prediction with Tree Based Model Classification. MECS Publisher.

ARNTZEN, H. and HVATTUM, L.M., 2020. Predicting match outcomes in association football using team ratings and player ratings. SAGE Publications.

BAATTITE, A., 2023. MACHINE LEARNING-BASED FOOTBALL TACTIC AND STYLE ANALYSIS.

GOES, F.R., MEERHOFF, L.A., BUENO, M.J.O., RODRIGUES, D.M., MOURA, F.A., BRINK, M.S., ELFERINK-GEMSER, M.T., KNOBBE, A.J., CUNHA, S.A., TORRES, R.S. and LEMMINK, K.A.P.M., 2020. Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. Wiley.

LAWRENCE, S. and CRAWFORD, G., 2021. Towards a digital football studies: current trends and future directions for football cultures research in the post-Covid-19 moment. Informa UK Limited.

PEÑA, J.L. and TOUCHETTE, H., 2012. A network theory analysis of football strategies.

PRETTO, F. and DE CASO, G., 2022. Development of a Football Analytics Web Application for Player Scouting.

SARMENTO, H., CLEMENTE, F.M., ARAÚJO, D., DAVIDS, K., MCROBERT, A. and FIGUEIREDO, A., 2017. What Performance Analysts Need to Know About Research Trends in Association Football (2012–2016): A Systematic Review. Springer Science and Business Media LLC.

SPENCER, B., HAWKEY, M. and ROBERTSON, S., Using contextual player movement and spatial control to analyse player passing trends in football.

TENGA, A.P.C., Reliability and Validity of Match Performance Analysis in Soccer.

ÜNSOY, O., 2022. DEVELOPING A DECISION-MAKING FRAMEWORK FOR PLAYER RECRUITMENT IN EUROPEAN FOOTBALL CLUBS.

WAKELAM, E., STEUBER, V. and WAKELAM, J., 2022. The collection, analysis and exploitation of footballer attributes: A systematic review. IOS Press.

WRIGHT, C., CARLING, C. and COLLINS, D., 2017. The wider context of performance analysis and it application in the football coaching process. Informa UK Limited.

BEAL, R., MIDDLETON, S.E., NORMAN, T.J. and RAMCHURN, S.D., 2021. Combining machine learning and human experts to predict match outcomes in

football: A baseline model, Proceedings of the AAAI Conference on Artificial Intelligence 2021, pp. 15447-15451.

CONSTANTINOU, A. and FENTON, N., 2017. Towards smart-data: Improving predictive accuracy in long-term football team performance. Knowledge-Based Systems, 124, pp. 93-104.

FLANAGAN, C.A., 2022. Stats Entertainment: The Legal and Regulatory Issues Arising from the Data Analytics Movement in Association Football. Part Two: Data Privacy, the Broader Legal Context, And Conclusions on the Legal Aspects of Data Analytics in Football. Entertainment and Sports Law Journal, 19, pp. 1.

GARNICA-CAPARRÓS, M. and MEMMERT, D., 2021. Understanding gender differences in professional European football through machine learning interpretability and match actions data. Scientific reports, 11(1), pp. 10805.

GÖLTAŞ, Y.T., 2023. Optimizing Football Lineup Selection Using Machine Learning. Optimizing Football Lineup Selection Using Machine Learning, .

GOUD, P.S.H.V., ROOPA, Y.M. and PADMAJA, B., 2019. Player performance analysis in sports: with fusion of machine learning and wearable technology, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) 2019, IEEE, pp. 600-603.

HERBERGER, T.A. and LITKE, C., 2021. The impact of big data and sports analytics on professional football: A systematic literature review. Digitalization, digital transformation and sustainability in the global economy: risks and opportunities, , pp. 147-171.

MEMMERT, D. and REIN, R., 2018. Match analysis, big data and tactics: current trends in elite soccer. German Journal of Sports Medicine/Deutsche Zeitschrift fur Sportmedizin, 69(3),.

PRATAS, J.M., VOLOSSOVITCH, A. and CARITA, A.I., 2017. PREDICTING KEY-GOAL SCORING IN FOOTBALL, BASED ON PERFORMANCE INDICATORS AND CONTEXTUAL FACTORS4. Analysis of goal scoring in football matches according to performance indicators and the context of competition, 1001, pp. 67.

RØNNINGEN, M.H., 2021. The genesis of data-driven decision-making in the world of soccer tactics: deciphering the potential of big data. The genesis of data-driven decision-making in the world of soccer tactics: deciphering the potential of big data, .

SARMENTO, H., MARCELINO, R., ANGUERA, M.T., CAMPANIÇO, J., MATOS, N. and LEITÃO, J.C., 2014. Match analysis in football: a systematic review. Journal of sports sciences, 32(20), pp. 1831-1843.

THAKKAR, P. and SHAH, M., 2021. An assessment of football through the lens of data science. Annals of Data Science, , pp. 1-14.

TUYLS, K., OMIDSHAFIEI, S., MULLER, P., WANG, Z., CONNOR, J., HENNES, D., GRAHAM, I., SPEARMAN, W., WASKETT, T. and STEEL, D., 2021. Game Plan: What AI can do for Football, and What Football can do for AI. Journal of Artificial Intelligence Research, 71, pp. 41-88.

A