Predicting Yellow and Red Cards

Candidate Number: BPSN2, Word Count: 10,207

Abstract

This research explores the predictive modeling of yellow and red card occurrences in football matches, utilizing a comprehensive dataset of historical match data. Two primary modeling approaches, Poisson regression and Negative Binomial regression, were employed to predict card events, with model selection based on AIC and BIC criteria. The study covers two distinct periods: pre-COVID and during COVID, providing insights into how external disruptions can impact card issuance.

The findings reveal that Poisson regression models based on AIC and BIC criteria consistently perform well in predicting both yellow and red cards, demonstrating their robustness across different phases of football history. Similarly, Negative Binomial regression models produce reliable predictions, further emphasizing their utility in handling over-dispersed count data.

Key factors influencing card occurrences include the country of the match, competition level, implied target goals, attendance, and kick-off time. The influence of these factors remained consistent over time, underscoring their significance in shaping card events. Additionally, the models exhibited resilience during the COVID-19 pandemic, suggesting that the fundamental dynamics of card issuance remained stable despite external disruptions.

This research contributes valuable insights into the factors driving card events in football and provides predictive models that can assist teams, referees, and stakeholders in anticipating and managing disciplinary actions. Furthermore, it highlights the need for future research to incorporate player-level data, real-time variables, and context-specific factors to enhance the accuracy and depth of card predictions. By addressing these aspects, the study advances our understanding of football match dynamics and offers a foundation for more sophisticated predictive modeling in the future.

Acknowledgement

Praise to God for giving us sufficient time as well as blessing us with the health and strength to complete this report. First and foremost, we would like to give our most sincere appreciation and heartfelt thanks to our supervisor, Dr. Niloufar Abourashchi for her patience, constant guidance, advice along with encouragement throughout the completion of the report. Her broad knowledge and constructive criticism have greatly assisted us in garnering new ideas as well as ushered us to constantly improve ourselves.

Special thanks also go to our friends and colleagues for their support and advice. Last but not least, a huge thanks to our families who gave full moral and financial support into completing this report

Contents

List of Figures

List of Tables

CHAPTER ONE INTRODUCTION

1.1 Background of Study

Football, also referred to as soccer internationally, has established itself as one of the most popular sports in the world, enthralling millions of spectators and generating significant revenue. In the modern football environment, player behaviour and relationships are intricately patterned and go well beyond just physical strength and skill. The distribution of yellow and red cards by match officials to punish transgressions and protect the integrity of the game is a major component within this complex framework. For teams, players, and fans, being able to predict when and why such disciplinary sanctions might be applied is crucial. In order to forecast when yellow and red cards will be issued during matches, this research digs into the field of football analytics. It does so by using sophisticated predictive modelling approaches.

A variety of elements play a part in modern football, from match circumstances and referee decisions to player abilities and tactical manoeuvres. Among these elements, the game's disciplinary aspect, particularly the issuing of yellow and red cards, stands out as a crucial component affecting the progress and result of a match. Understanding the patterns and determinants of card issuance can provide important insights into the complex game dynamics, player behaviour, and referee decision-making procedures. The goal of this study is to use data science approaches to estimate the possibility of issuing yellow and red cards during football games.

Football referees employ yellow and red cards as necessary disciplinary tools to control player behaviour and maintain the integrity of the game. A yellow card serves as a warning, indicating that a player's behaviour is not acceptable without instantly calling for their expulsion. In contrast, a red card requires that a player be removed from the field, often due to severe misconduct or the accumulation of two yellow cards.

The effects of card distribution go far beyond specific players and teams, having a significant impact on the dynamics of the entire match. A team's chances of success are hampered by the numerical disadvantage that red cards cause. The accumulation of yellow cards can also result in player suspensions, which will affect the team's game plan for consecutive games.

As football continues to evolve, the intricate interplay between players, referees, and the rules of the game remains a captivating aspect for both enthusiasts and analysts alike. The quest to understand and predict card issuance patterns involves not only the collection of vast amounts of data but also the utilization of cutting-edge data science methodologies.

Intriguingly, the consequences of yellow and red cards transcend individual matches. The accumulation of cards throughout a season can lead to suspensions that significantly impact team strategies. The realm of football analytics,

specifically the prediction of yellow and red card issuance, represents a captivating intersection of sports, data science, and strategy. This paper will explore the mysteries of football's disciplinary facet, providing valuable insights for teams, players, and fans alike.

1.2 Significance and Research Objectives

This research carries significance not solely within sports analysis, but it also transcends to broader discussions regarding data-driven decision-making. By delving into football analytics and constructing predictive models for card issuance, this study contributes to the evolving arena of sports analytics literature. Furthermore, it serves as a testament to the transformative power of data-driven insights in shaping competitive environments.

It's noteworthy to incorporate the impact of the COVID-19 pandemic on football dynamics. The dataset used in this research extends until 2021, capturing a period when the pandemic disrupted normal operations. A distinctive consequence was the suspension of matches and altered attendance dynamics. Beginning in mid-April 2020, attendance figures dwindled to zero, reflecting the pandemic's far-reaching implications on football's landscape.

The primary goal of this dissertation is to construct robust predictive models capable of estimating the likelihood of yellow and red cards being issued during a football match. To achieve this, Poisson regression and negative binomial regression, two statistical techniques well-suited for count data were employed. The dataset used for this analysis encompasses various features, including country, competition level, kick-off time, referee details, implied bookmaker odds (supplied and total goals), attendance, and referee experience level. The aim is to determine which factors significantly influence the issuance of cards and how these models can enhance our understanding of refereeing decisions in football.

Subsequent chapters will expound on the literature review, data preprocessing, feature engineering, methodology, results, and the conclusion of this research endeavor.

CHAPTER TWO LITERATURE REVIEW

This chapter provides a summary of the relevant studies, methodologies, and literature that were employed in the following chapters.

2.1 Related Studies on Football Data

A previous paper focused on modeling the in-play football betting market by applying financial mathematical concepts and machine learning techniques (Divos, 2020). It aimed to value and manage risks of in-play football bets, utilizing high-frequency data sets and the direct link between game fundamentals and betting market outcomes. The paper introduced the application of financial derivatives' results, such as risk-neutral measures and arbitrage-freeness, to in-play football betting through a Poisson process-based model analogous to the Black-Scholes model. It also proposes a Local Intensity model to address the implied intensity smile observed in football betting. The paper presents a Microscopic Model that incorporates variables like ball position and team possession, arriving at a simplified model with predictive capabilities for short time intervals. Additionally, a First Half Indicators Model is introduced, utilizing machine learning to predict second-half goal intensities based on first-half indicators. The study finds that the Microscopic Model outperforms the First Half Indicators Model for short delays, due to the relevance of initial ball position and team possession within the first 30 seconds.

Decroos et al. (2019) set out to redefine the assessment of player contributions in soccer by going beyond traditional goal-based metrics. Their objective was to emphasize the significance of various player actions that influence match outcomes. Employing a data-driven approach, they meticulously collected and analyzed player movement data, tracking actions such as passes, tackles, and positioning. The study aimed to create a comprehensive framework for valuing these actions and predicting their impact on match dynamics. Through advanced predictive modeling techniques, the researchers sought to elucidate the nuanced ways in which players contribute to the ebb and flow of a game beyond scoring goals. Their findings revealed that accounting for a wider array of player actions enhances the accuracy of predicting match outcomes and offers a more holistic understanding of player influence.

Peña et al. (2012) undertook an investigation into the intricate factors influencing the issuance of yellow and red cards in football matches. Their primary objective was to uncover potential biases or discriminatory practices in referee decisions. Employing a rigorous statistical analysis, the researchers examined a large dataset of matches, focusing on the relationship between player behaviors, match situations, and card allocations. Through a comprehensive examination of these variables, the study aimed to discern patterns that shed light on the referees' decision-making process. The findings of their research highlighted that certain player profiles, along with the stage of the match and the overall

context, significantly influenced the likelihood of receiving a red card. This illuminated the presence of implicit biases in card issuance, drawing attention to the complexities of referee judgment and its potential impact on match outcomes.

Kovalchik et al. (2016) contributed to the burgeoning field of predictive analytics in football by investigating the factors associated with yellow card accumulation. The primary objective of their research was to identify specific risk factors that lead to players receiving yellow cards. Leveraging player movement data from matches, the researchers utilized machine learning techniques to uncover underlying patterns. By training predictive models on historical data, they aimed to pinpoint the combination of player actions, positioning, and match situations that contribute to disciplinary actions. The results of their study revealed key predictors that could forecast instances of yellow card issuance. This pioneering research showcased the potential of advanced analytics in anticipating player behavior and provided a foundation for further explorations into predicting disciplinary outcomes.

Clarke & Norman (2019) embarked on an exploration of the integration of data science methodologies in sports analytics, specifically within the context of football. Their objective was to underscore the transformative impact of data analytics on various aspects of the sport. Through a comprehensive literature review and analysis of existing research, the study aimed to elucidate the range of applications enabled by data-driven insights. From enhancing player performance analysis to refining strategic decision-making, the researchers highlighted the multifaceted ways in which data science contributes to modern sports practices. Their findings underscored the growing significance of data analytics in unlocking hidden patterns, optimizing training regimes, and shaping the trajectory of competitive sports.

Borland & MacDonald (2020) delved into the unprecedented impact of the COVID-19 pandemic on the dynamics of football matches, particularly the absence of fan attendance. Their objective was to examine the ramifications of playing in empty stadiums and the subsequent alteration in the atmosphere of matches. Through a combination of qualitative and quantitative analysis, the researchers assessed changes in team dynamics, player behavior, and overall match environment. By comparing pre-pandemic matches with those played during the pandemic, they aimed to capture the qualitative and quantitative shifts in game dynamics. The findings of their study revealed a notable shift in the emotional and psychological dynamics of matches due to the lack of fan engagement. This demonstrated the intricate connection between fan presence and match outcomes, underlining the crucial role of spectator influence in the football ecosystem.

Fernandez-Corugedo & McMahon (2021) delved into the economic implications of the COVID-19 pandemic on the market for football players. Their primary objective was to understand how the disruptions caused by the pandemic affected player transfers, contract negotiations, and the overall football market. Through comprehensive data analysis and economic modeling, the researchers

examined shifts in player valuations, market activity, and financial trends. By analyzing transfer data, contract extensions, and market dynamics during the pandemic, they aimed to provide insights into the impact on player movement and financial decisions. The findings of their research highlighted the altered landscape of player valuations and the subsequent adjustments in market transactions. This demonstrated the far-reaching consequences of the pandemic on football's financial ecosystem, emphasizing its complex interplay with on-field dynamics.

2.2 Related Studies on Modelling Football Data

Karlis & Ntzoufras (2003) aimed to extend the application of Poisson regression models to the analysis of sports data by introducing bivariate Poisson models. Their primary objective was to develop a methodology for modeling the joint distribution of goals scored by two competing teams in a match, incorporating dependencies between the teams' scoring behaviors. Through this approach, they sought to enhance the predictive accuracy of goal outcomes and provide deeper insights into the dynamics of sports events. The researchers collected historical football match data, including goals scored by both teams, and utilized the bivariate Poisson models to capture the goal-scoring interactions. Their findings indicated that the bivariate Poisson models outperformed traditional Poisson models by accounting for the interdependence between teams' goal-scoring behaviors. This research demonstrated the efficacy of extending Poisson regression to better capture the complexities of sports events and improve predictive accuracy.

Dixon & Coles (1997) focused on modeling association football scores and exploring inefficiencies in the football betting market using the Poisson regression framework. The researchers' objective was to develop a statistical model that could predict the distribution of goals scored by teams in football matches. They employed a Poisson regression approach to model the number of goals scored by each team based on historical match data. By incorporating various covariates such as team strength, home advantage, and opposition quality, they aimed to improve the accuracy of score predictions. The findings of their study revealed that the Poisson regression model provided a solid foundation for predicting match outcomes and identifying potential discrepancies in the football betting market. This research highlighted the significance of statistical modeling in understanding goal-scoring patterns and uncovering market inefficiencies in sports betting.

Mohan & Samuelsen (2017) directed their research toward modeling the number of goals scored by a single team in a football match using negative binomial regression. Their primary objective was to develop a methodology that could capture the overdispersion often present in goal-scoring data. By incorporating team-specific covariates such as offensive and defensive abilities, they aimed to provide a robust model for predicting team goal outcomes. The researchers

collected data from historical football matches and employed negative binomial regression to account for the variance in goal-scoring rates. Their findings indicated that the negative binomial model effectively addressed overdispersion and yielded more accurate predictions compared to traditional Poisson models. This research showcased the relevance of negative binomial regression in accounting for the variability inherent in goal-scoring data, enhancing the precision of match outcome predictions.

In addition to its application in sports analytics, the utilization of negative binomial regression has been widely recognized in healthcare research to examine factors influencing various healthcare outcomes. Islam et al. (2020) utilized the negative binomial regression model to investigate triggering factors associated with the utilization of antenatal care visits in Bangladesh. By incorporating this statistical technique, the researchers were able to account for overdispersion and effectively analyze count data related to healthcare visits. This study highlights the versatility of negative binomial regression beyond the realm of sports analytics and underscores its utility in studying complex real-world phenomena across diverse disciplines.

In the study conducted by Baio & Blangiardo (2010), the researchers introduced a Bayesian hierarchical modeling approach to predict football match outcomes. The primary objective was to develop a model capable of providing probabilistic forecasts for football match results. The methodology involved a hierarchical structure that incorporated various levels of information, including team-specific attributes, historical performance, and match-specific covariates. Additionally, they integrated the concept of team strength, which evolved over time, allowing the model to adapt to changing team dynamics. The findings of the study were promising, as the proposed Bayesian hierarchical model demonstrated superior predictive accuracy compared to traditional models. Moreover, it provided uncertainty estimates for match outcomes, which is valuable for both sports analysts and betting enthusiasts. The research highlighted the importance of considering team-specific characteristics and temporal effects in football prediction models, and it showcased the advantages of a Bayesian framework in this context. This study contributes to the growing field of football analytics by offering a robust modeling approach that can enhance the accuracy of match result predictions and aid decision-making in various football-related domains.

CHAPTER THREE METHODOLOGY

3.1 Data

This section provides a comprehensive overview of the variables present in the data set used for predicting yellow and red cards in football matches. The data set encompasses a diverse range of factors that contribute to the number of yellow and red cards.

The data set used in this study is obtained by Smartodds, a private consultancy that provides services for professional gamblers. It comprises a comprehensive collection of football matches from multiple European leagues across various seasons, countries, and competition levels. Each data point is a unique football match, capturing a wide range of attributes that offer insights into match dynamics, disciplinary actions, and contextual factors. These data contain 27255 observations from August 2014 until December 2021. The data set includes the following variables (the data type is checked using the glimpse() function in R):

Table 1: Variable Description

Column Name (Data Type)	Description
country (character)	Designates the geographical location of each match, identifying the country where the football league operates. This categorical variable helps contextualize the matches within their respective regions.
competition_level (double)	Classifies matches based on the level of competition they belong to, such as top-tier leagues or lower divisions. This categorical variable enables the examination of disciplinary trends across various competitive environments.
kick_off_datetime (date and time)	Provides precise temporal information for each match
referee (character)	Records the name of the referee overseeing each match.
sup_implied, tg_implied (double)	Provide implied odds from the betting market for match outcomes and total goals based on the available Asian handicap
team1_yc, team2_yc, team1_rc, team2_rc (double)	Quantify the number of yellow and red cards received by each team.

Column Name (Data Type)	Description
attendance_value (double)	Quantifies the number of spectators present during a match.

Asian handicap is a betting term that originates in Asia and is primarily applied to football (Isaiah, 2023). A handicap indicates that one team effectively leads the other. Due to Asian handicaps, there are only two outcomes that can occur in a game as it eliminate the chance of a draw. The favoured team that is at a disadvantage is denoted with a minus sign (-) while a plus sign (+) is used to indicate the handicap advantage enjoyed by the underdog team.

3.2 Statistical Model

3.2.1 Poisson Regression Model

Poisson regression is a statistical method used to model count data, such as the number of goals scored in a football match or the number of yellow cards received by players (Agresti, 2006). The Poisson distribution is commonly employed to capture the probability of rare events occurring in a fixed interval. In the context of football analytics, Poisson regression models can be used to predict the frequency of certain discrete events, providing insights into match dynamics.

The Poisson distribution serves as the foundation for Poisson regression. The probability mass function (PMF) of the Poisson distribution is given by:

$$P(Y = y) = \frac{e^{-\lambda} \cdot \lambda^y}{y!}$$

Where:

- y is the count of events.
- λ is the average rate of events.

In the context of Poisson regression, the model estimates the expected count λ based on predictor variables. The general equation for Poisson regression is:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

Where:

- λ_i: The expected count of the event for observation i.
- x_{ji}: The value of predictor j for observation i.

- β_j: The coefficient associated with predictor j.
- p: The number of predictors in the model.

The log-link function ensures that the predicted values remain positive. Poisson regression assumes that the mean and variance of the response variable are equal.

This modeling approach has been applied successfully in football analytics. For instance, Karlis & Ntzoufras (2003) analyzed sports data using bivariate Poisson models to predict joint goal distributions. Additionally, Dixon & Coles (1997) utilized Poisson regression to model association football scores and identify inefficiencies in the football betting market.

3.2.2 Negative Binomial Regression

Negative binomial regression is an extension of the Poisson regression model that accounts for overdispersion (Agresti, 2006), which occurs when the variance of the response variable is greater than its mean. Overdispersion is common in count data, where the assumption of equal mean and variance may not hold. Negative binomial regression is particularly useful when dealing with rare events or data with excessive variation.

The negative binomial distribution accommodates overdispersion. The probability mass function (PMF) of the negative binomial distribution is:

$$P(Y = y) = \frac{\Gamma(y + r)}{y! \cdot \Gamma(r)} \cdot (1 - p)^r \cdot p^y$$

Where:

- y is the count of events.
- r is the shape parameter that controls dispersion.
- p is the probability of success.

In negative binomial regression, the model equation becomes:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \log(\theta)$$

Where all variables have the same meanings as in the Poisson regression equation.

The negative binomial regression model allows for greater flexibility in handling overdispersed count data, making it a suitable choice for predicting football-related events that may exhibit varying levels of variance.

This methodology has been employed in football research. Mohan & Samuelsen (2017) utilized negative binomial regression to model the number of goals scored by a team in a football match, accounting for overdispersion and capturing the complexities of goal-scoring patterns. Furthermore, while not exclusively about negative binomial regression, Clarke & Norman (2019) discuss the role of data analytics in sports, including football, which provides insights relevant to understanding the broader application of advanced regression techniques in sports analytics.

3.2.3 Akaike Information Criterion (AIC)

AIC is a widely-used statistical measure used for model selection and comparison. It quantifies the trade-off between model fit and model complexity. It plays a crucial role in determining the optimal negative binomial regression model of analysis.

AIC is defined as:

$$AIC = -2 \cdot \ln(L) + 2k$$

Where:

- L represents the likelihood of the model, which measures how well the model fits the data.
- k is the number of estimated parameters in the model.

A lower AIC value indicates a better balance between model fit and complexity. Therefore, when comparing different models, the one with the lowest AIC is generally preferred.

3.2.4 Bayesian Information Criterion (BIC)

BIC, like AIC, is used for model selection and evaluation. It also addresses the trade-off between model fit and complexity but places a stronger penalty on models with more parameters, which helps prevent overfitting.

BIC is defined as:

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$$

Where:

- L is the likelihood of the model.
- k is the number of estimated parameters.
- n is the sample size.

Similar to AIC, a lower BIC value indicates a better-fitting model. However, BIC tends to be more conservative in selecting simpler models when compared to AIC because it imposes a larger penalty for additional parameters.

3.2.5 Accuracy Checking for Predicted Values

The accuracy checking of the predicted values were measured using the Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). RMSE measures the average magnitude of the errors between the actual and predicted values. It is commonly used to assess the overall accuracy of a model's predictions. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

Where: - e_t represents the forecast error at time t, calculated as $e_t = y_t - \hat{y}_t$.

- y_t is the actual observed value at time t.
- $\hat{y}_t$ is the fitted (forecasted) value at time t.
- n is the number of values being forecasted.

MAE measures the average magnitude of the errors between the actual and predicted values. It is less sensitive to outliers compared to RMSE. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

Where: - e_t represents the forecast error at time t, calculated as $e_t = y_t - \hat{y}_t$.

- y_t is the actual observed value at time t.
- $\hat{y}_t$ is the fitted (forecasted) value at time t.
- n is the number of values being forecasted.

CHAPTER FOUR DATA EXPLORATION

4.1 Data Quality Assessment

This chapter provides a thorough investigation of the data quality. Any empirical investigation must consider data quality since the correctness and completeness of the underlying data are crucial to the validity and dependability of research conclusions. This chapter delves into the methods used to assess the data set for duplicates and missing values.

4.1.1 Duplicate Data Detection

Prior to beginning any research, it is crucial to guarantee the data set's integrity by locating and removing any potential duplicate items. Duplicate data can seriously reduce the accuracy of results, skewing conclusions and causing erroneous decisions. The analysis revealed that the data set under investigation was free from any duplicate records.

4.1.2 Missing Data Analysis

Another serious issue that has the potential to skew the results of statistical analysis is missing data. Inaccurate or incomplete data might provide biased findings and false conclusions. The data set's character and numeric variables were meticulously examined for missing values in order to remedy this issue. Notably, the numeric columns of the variables "sup_implied" and "tg_implied" consistently showed six missing values for each of them. Additionally, the character variable "referee" showed 359 instances of significant missing values. Identification of these missing values is a crucial first step in the development of the imputation techniques that will be used in later chapters to maintain the completeness of the data set and the accuracy of the analysis.

4.1.3 Unique Values of Categorical Variables

As part of the data quality assessment, a meticulous examination of the unique values within the "country" and "referee" variables was conducted. The variable "country" encompasses a total of five unique values, namely "France," "England," "Germany," "Spain," and "Italy." This succinct distribution underscores the international representation of football matches within the dataset, encapsulating matches from some of the most prominent European footballing nations.

In contrast, the "referee" variable exhibits a greater diversity, comprising 387 unique values, including instances of missing data (NA). This wide array of unique referee names reflects the diversity of match officials that oversee football

matches. While this diversity introduces a level of complexity, it also contributes to the richness and authenticity of the dataset. The presence of missing values in the "referee" variable prompts further consideration regarding the handling of these instances, a topic that will be explored in subsequent sections of this study.

The results of the data quality assessment show how meticulously the research results were preserved as credible. The data set's originality is confirmed by the lack of duplicate entries, while the detection of missing values provides the opportunity for informed imputation techniques. By proving the data set's dependability, this chapter establishes the foundation for later investigations.

On top of this foundation, advanced modelling approaches, inferential statistics, and exploratory data analysis will be undertaken in the chapters that follow.

4.2 Data Transformation and Imputation

The data set underwent transformation and imputation procedures after being thoroughly evaluated for data quality in order to improve its analytical usefulness. The processes taken to reconstruct the data and deal with missing values are described in this chapter, ensuring that the subsequent analyses are comprehensive and correct.

4.2.1 Missing Data Imputation

Addressing the presence of missing data in the "referee" variable required a systematic approach that considered the unique characteristics of each country. To begin, an examination of the number of missing referees by country was conducted. The distribution revealed that Germany had a particularly high count of missing values (316), whereas other countries exhibited relatively fewer instances of missing referees.

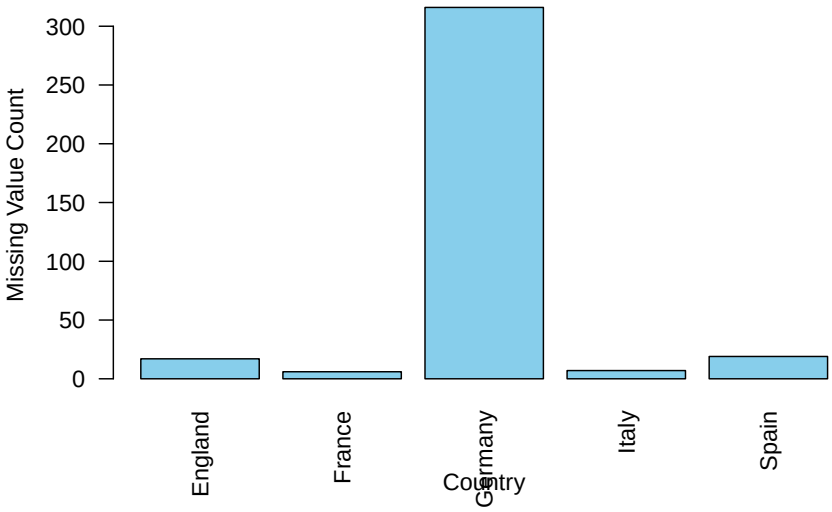

Figure 1: Missing Referee by Country

To address this, a mode analysis was employed, focusing on identifying the most common referee name within each country. The mode serves as a representative value that aligns with the prevailing trend. For instance, in England, "Anthony Taylor" emerged as the mode referee, while "Ruddy Buquet" held this distinction in France. Intriguingly, the mode for Germany was also a missing value, which prompted further exploration.

Table 2: Mode of Referee by Country

Country	Mode
England	Anthony Taylor
France	Ruddy Buquet
Germany	NA
Italy	Daniele Doveri
Spain	Arcerdiano Monescillo

Recognizing the need for a reliable imputation strategy for Germany, the second mode was examined, revealing "Felix Brych" as the runner-up in terms of frequency. Armed with this insight, the missing referee entries for Germany were imputed based on this second mode.

By tailoring the imputation approach to the specific characteristics of each country, this procedure ensures that the "referee" variable retains its contextual accu-

racy. The imputed values based on mode serve as credible approximations that preserve the integrity of the data. This meticulous handling of missing data contributes to the reliability of subsequent analyses and enhances the overall quality of the dataset.

To determine the appropriate imputation method for handling missing values in the "sup_implied" and "tg_implied" variables, a careful analysis of their summary statistics was conducted. For the "sup_implied" variable, the mean (0.31282) closely aligns with the median (0.27508), indicating a potential normal distribution of data.

Table 3: Summary Statistics of sup_implied

Mininmum	First Quartile	Median	Mean	Third Quartile	Maximum
-3.40325	-0.07054	0.27508	0.31282	0.66181	4.21039

This is further proven by the following distribution plot:

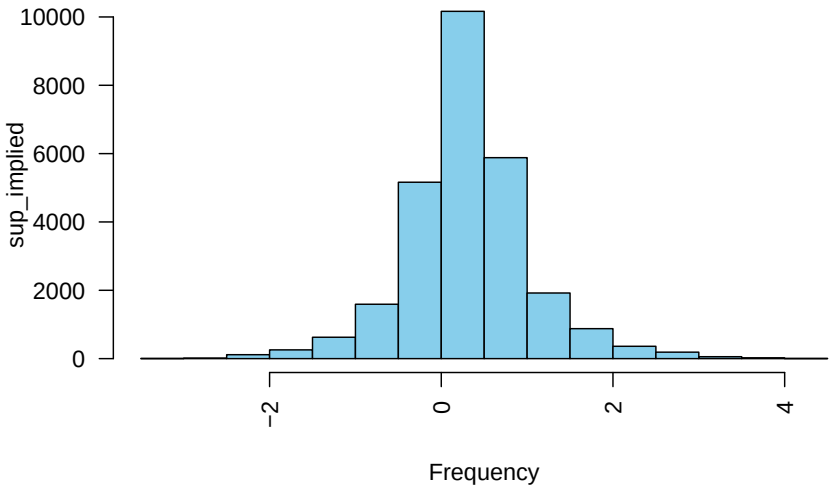

Figure 2: Distribution of sup_implied

Turning to the "tg_implied" variable, the mean (2.603) exhibits a proximity to the median (2.536), suggesting that the data distribution could possess a reasonable level of symmetry.

Table 4: Summary Statistics of tg_implied

Mininmum	First Quartile	Median	Mean	Third Quartile	Maximum
1.599	2.325	2.536	2.603	2.798	5.267

This can be seen in the following plot:

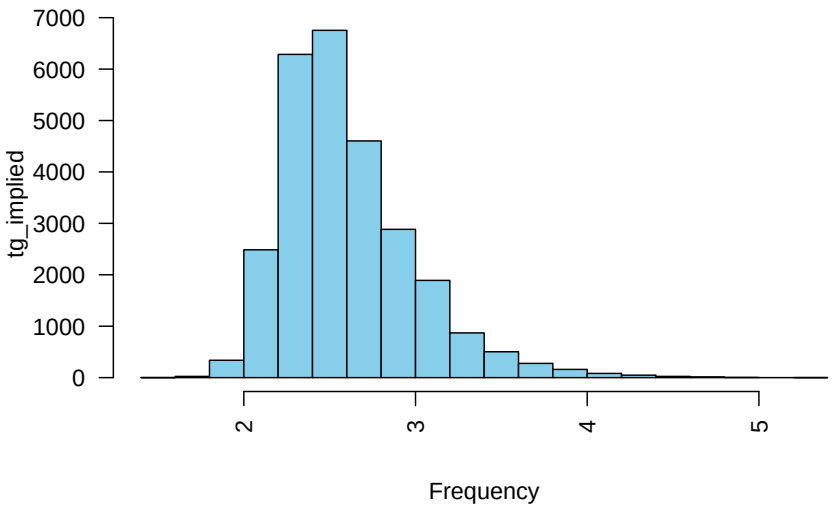

Figure 3: Distribution of tg_implied

Mean imputation involves replacing missing values with the average value of the non-missing observations within the same variable. This approach capitalizes on the symmetric nature of the distributions and the relatively consistent central tendency of the variables. By imputing missing values with the mean, the resulting dataset maintains the general characteristics of the original distribution, thus preserving the underlying structure of the data.

This imputation strategy is well-suited for continuous variables like "sup_implied" and "tg_implied," where the mean serves as a representative value that aligns with the overall trend of the data. The decision to opt for mean imputation is supported by the absence of influential outliers and the distribution's relative symmetry. By consistently applying mean imputation, the dataset is enriched with imputed values that enhance the completeness of the variables while adhering to the statistical attributes of the data.

4.2.2 Data Transformation

The original column "kick_off_datetime" was divided into the two separate columns "kick_off_date" and "kick_off_time" to increase the data set's level of granularity. This division makes it possible to examine match scheduling in greater detail. Further classification of the "kick_off_time" column allowed matches to be labelled as occurring in the afternoon (1) if they occur between

10am and 4pm or the evening (2) if they occur between 4pm and 10pm, respectively. This subtle categorization captures the temporal fluctuations in match results, which helps to provide a more accurate study of match dynamics during the course of the day.

The columns "team1_yc" and "team2_yc" were combined into a single "total_yc" column to capture the total amount of yellow cards that both teams have been issued in a match in the search for more illuminating variables. Similar to this, the data from the "team1_rc" and "team2_rc" columns was merged to create the "total_rc" column, which tallied the red card statistics for both teams. In addition to streamlining the data set, these changes produce variables that provide a more thorough summary of match-related sanctions.

To improve the usability of the dataset, two pivotal columns, "matches" and "referee_group," have been seamlessly integrated. These additions introduce crucial dimensions that illuminate the world of football officiating. The "matches" column quantifies the number of matches each referee has overseen, offering a window into their tenure and familiarity with the duties of a match official. This metric lays the groundwork for categorizing referees into distinct groups based on their level of experience.

Furthermore, the "referee_group" column ushers referees into four distinguished categories: Novice, Experienced, Veteran, and Elite. This categorization springs from a nuanced approach that meticulously considers the quartiles of matches officiated. Novice referees, proudly donning the "Novice" badge, are those who have marshaled fewer than 18 matches. They stand at the inception of their officiating journey, gradually accumulating invaluable experience. "Experienced" referees, presiding over a modest range of 18 to 54 matches, occupy the middle ground, where their skills and familiarity with the role progress steadily. The "Veteran" cadre encompasses referees with a match count spanning from 55 to 120, marking a rich repository of experience and a substantial body of work. Finally, the "Elite" echelon embraces referees who have graced more than 120 matches, epitomizing a level of expertise and tenure that commands reverence.

Table 5: Summary Statstics of Matches per Referee

Mean	First Quartile	Median	Third Quartile	Standard Deviation	Minimum	Maximum
70.42636	18	54	120	61.21496	1	482

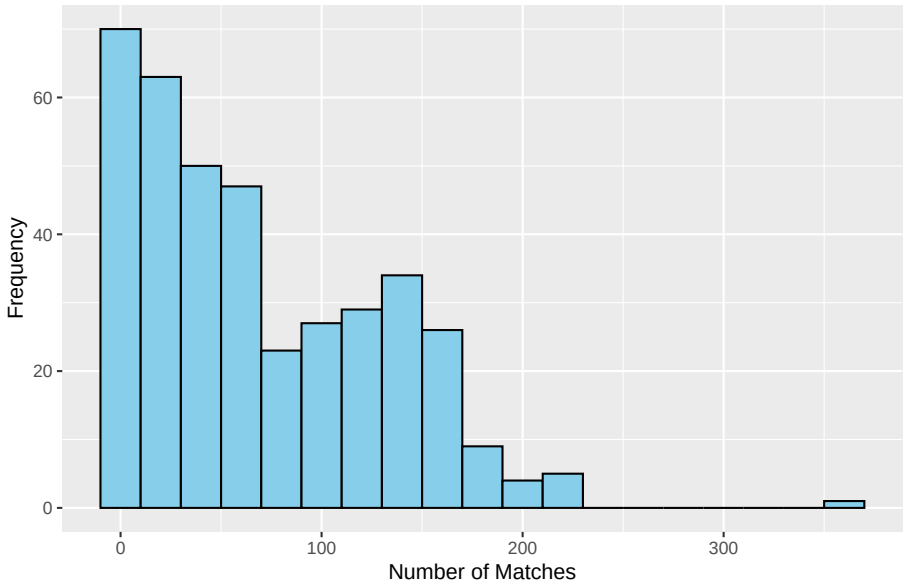

Figure 4: Distribution of Matches per Referee

The data set has been improved by these data transformations and imputations to a state that is suitable for insightful analyses. These improvements will be utilised in the following chapters to explore exploratory data analysis, inferential statistics, and predictive modelling. This chapter ensures that the data set is ready for robust analysis by resolving missing data and adding new variables, thereby enhancing the depth and rigour of the research findings.

CHAPTER FIVE RESULTS AND DISCUSSION

Further exploration unveiled a significant inflection point in the dataset. Starting April 2020, a noticeable drop in attendance figures emerged, a testament to the far-reaching impact of the COVID-19 pandemic on footballing spectacles. This pivotal observation prompted the dataset being divided into two distinct portions; the pre-COVID era and the COVID era, recognizing the shifts in football dynamics influenced by the pandemic. The pre-COVID data consists of 20879 observations and the COVID era has 6376 observations.

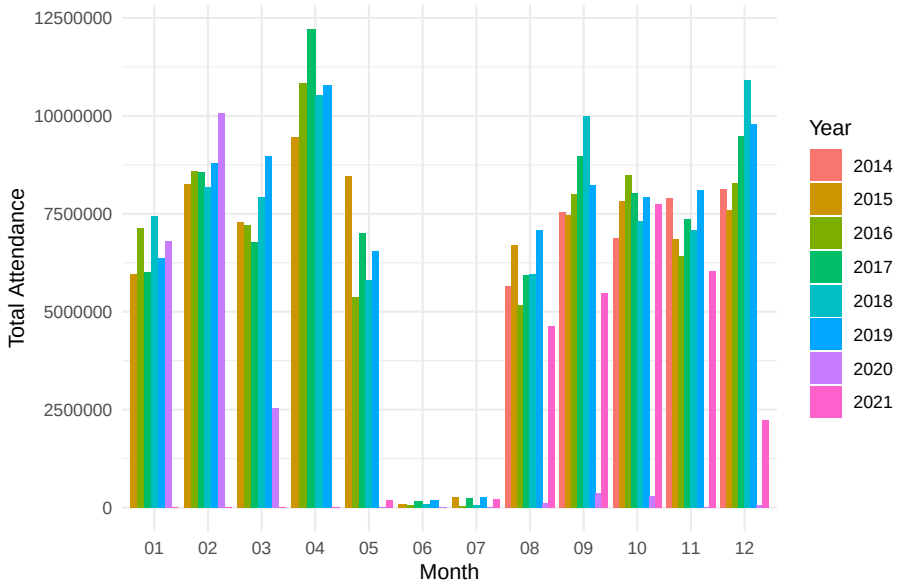

Figure 5: Distribution Attendees Across Month and Year

Within the pre-COVID era, 70% of the data is allocated to the training set and 30% for validation. The training includes 14616 lines of data while the validation set consists of 6263 observations. This division ensures the predictive models have the precision needed to navigate the intricacies of football dynamics before the pandemic's influence reshaped the landscape.

5.1 Modelling Yellow Cards

In this chapter, the results of predictive modeling for yellow cards in football matches using Poisson regression and negative binomial regression are presented.

Two models for each regression method are produced, each selected based on different information criteria: AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). The objective was to determine which model provided a better fit for predicting the total number of yellow cards, and the performance was evaluated using key metrics, including RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error).

5.1.1 Poisson Regression

The models were initially trained using a dataset that included several predictors, such as country, competition level, implied support, implied target goals, attendance value, kick-off time, matches, and referee group. The AIC-based model produced from Stepwise forward and backward regression retained more predictors in the final model, including country, competition level, implied support, implied target goals, attendance value, kick-off time, matches, and referee group.

Table 6: Summary of Best Model Based on Lowest AIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	1.516e+00	4.237e-02	35.768	$<$ 2e-16
countryFrance	8.240e-02	1.524e-02	5.407	6.40e-08
countryGermany	1.544e-01	1.392e-02	11.093	$<$ 2e-16
countryItaly	3.630e-01	1.601e-02	22.674	$<$ 2e-16
countrySpain	4.540e-01	1.340e-02	33.891	$<$ 2e-16
competition_levelL2	2.240e-02	1.091e-02	2.053	0.040076
sup_implied	-3.727e-02	6.345e-03	-5.874	4.25e-09
tg_implied	-1.454e-01	1.321e-02	-11.008	$<$ 2e-16
attendance_value	1.504e-06	3.734e-07	4.027	5.66e-05
kick_off_timecatEvening	-4.652e-02	9.039e-03	-5.147	2.65e-07
matches	4.229e-04	1.167e-04	3.623	0.000291
referee_groupExperienced	7.565e-02	1.902e-02	3.978	6.95e-05
referee_groupNovice	2.109e-02	3.326e-02	0.634	0.526078
referee_groupVeteran	1.108e-02	1.234e-02	0.898	0.369336

In contrast, the BIC-based model selected a simpler model with fewer predictors, including country, implied support, implied target goals, attendance value, and kick-off time.

Table 7: Summary of Best Model Based on Lowest BIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	1.621e+00	3.285e-02	49.342	$<$ 2e-16
countryFrance	6.381e-02	1.429e-02	4.466	7.95e-06
countryGermany	1.544e-01	1.356e-02	11.387	$<$ 2e-16
countryItaly	3.419e-01	1.387e-02	24.655	$<$ 2e-16
countrySpain	4.353e-01	1.252e-02	34.758	$<$ 2e-16
sup_implied	-3.448e-02	6.148e-03	-5.608	2.05e-08
tg_implied	-1.503e-01	1.308e-02	-11.493	$<$ 2e-16
attendance_value	1.164e-06	3.131e-07	3.716	0.000202
kick_off_timecat-Evening	-4.506e-02	9.016e-03	-4.998	5.80e-07

To assess the predictive accuracy of the two models, the performance metrics were calculated.

Table 8: Model Performance of Poisson Regression

Model	RMSE	MAE
AIC	1.973	1.585
BIC	1.974	1.585

Interestingly, both models displayed remarkably similar performance, with the AIC-based model yielding a slightly lower RMSE of approximately 1.973 and an MAE of about 1.585. The BIC-based model produced a very similar RMSE of approximately 1.974 and an MAE of about 1.585. This suggests that, despite the differences in model complexity and predictor selection, both models perform almost equally well in predicting the total number of yellow cards in football matches.

The similarity in performance between the AIC-based and BIC-based models raises some interesting questions about model complexity and parsimony. The AIC, which tends to favor more complex models, suggested the inclusion of additional predictors, such as competition level and referee group. In contrast, the

BIC, which penalizes model complexity, led to a simpler model with fewer predictors. However, despite these differences in model composition, both models achieved virtually identical predictive accuracy.

In interpreting the coefficients of the Poisson regression models, it's crucial to understand their implications for yellow card issuance during football matches. The coefficients provide insights into how each predictor influences the likelihood of a player receiving a yellow card. For instance, positive coefficients, such as those associated with 'countrySpain' and 'countryItaly,' suggest an increased likelihood of yellow card issuance when these countries are involved in matches. Conversely, negative coefficients, such as 'sup_implied' and 'tg_implied,' indicate that higher implied bookmaker odds for support and target goals are associated with a reduced likelihood of yellow cards. Additionally, the 'attendance_value' coefficient signifies that higher attendance figures correlate with a slightly elevated probability of yellow cards. While these interpretations provide valuable insights, it's essential to acknowledge the assumptions underlying the models. Assumptions related to the Poisson distribution, such as constant variance and independence of events, were considered during model construction. Any deviations from these assumptions, if present, were addressed through appropriate data transformations. Furthermore, a rigorous comparison between the AIC-based and BIC-based models was conducted, involving a likelihood ratio test. Surprisingly, the models exhibited strikingly similar performance metrics, implying that their predictive accuracy is nearly identical. This suggests that despite differences in model complexity and predictor selection, both models effectively capture the underlying patterns of yellow card issuance.

5.1.2 Negative Binomial Regression

In this section, the results of negative binomial regression models for predicting yellow cards in football matches are explored. Two models were applied, one based on AIC and another based on BIC, to understand how different model selection criteria can impact the predictive performance of the models. The negative binomial regression was chosen due to its ability to handle overdispersed count data, which is often the case with yellow card occurrences in football.

Similar to the Poisson regression models, the negative binomial regression models was initially trained with a set of predictors, including country, competition level, implied support, implied target goals, attendance value, and kick-off time, along with the number of matches and referee group. The AIC-based model retained all these predictors in the final model.

Table 9: Summary of Best Model Based on Lowest AIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	1.537e+00	4.615e-02	33.301	$<$ 2e-16
countryFrance	8.240e-02	1.524e-02	5.407	6.42e-08
countryGermany	1.544e-01	1.392e-02	11.092	$<$ 2e-16
countryItaly	3.630e-01	1.601e-02	22.672	$<$ 2e-16
countrySpain	4.540e-01	1.340e-02	33.888	$<$ 2e-16
competition_ levelL2	2.240e-02	1.091e-02	2.053	0.040085
sup_implied	-3.727e-02	6.345e-03	-5.874	4.26e-09
tg_implied	-1.454e-01	1.321e-02	-11.007	$<$ 2e-16
attendance_ value	1.503e-06	3.734e-07	4.026	5.67e-05
kick_off_ timecat- Evening	-4.652e-02	9.039e-03	-5.147	2.65e-07
matches	4.229e-04	1.167e-04	3.623	0.000291
referee_group Elite	-2.108e-02	3.326e-02	-0.634	0.526131
referee_group Experienced	5.457e-02	3.022e-02	1.806	0.070923
referee_group Veteran	-1.001e-02	3.029e-02	-0.330	0.741085

While the BIC-based model selected a simplified version of the model, excluding competition level and referee group.

Table 10: Summary of Best Model Based on Lowest BIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	1.621e+00	3.285e-02	49.339	$<$ 2e-16
countryFrance	6.381e-02	1.429e-02	4.466	7.96e-06
countryGerman y	1.544e-01	1.356e-02	11.386	$<$ 2e-16
countryItaly	3.419e-01	1.387e-02	24.653	$<$ 2e-16
countrySpain	4.353e-01	1.252e-02	34.756	$<$ 2e-16
sup_implied	-3.448e-02	6.148e-03	-5.607	2.05e-08
tg_implied	-1.503e-01	1.308e-02	-11.493	$<$ 2e-16
attendance_ value	1.164e-06	3.132e-07	3.716	0.000202

Coefficients	Estimate	Standard Error	z value	p-value
kick_off_timecatEvening	-4.506e-02	9.016e-03	-4.997	5.81e-07

The coefficients derived from both the AIC-based and BIC-based negative binomial regression models offer valuable insights into the factors influencing the issuance of yellow cards in football matches. Notably, the coefficients for country indicators, including France, Germany, Italy, and Spain, reveal the substantial impact of the country in which the match is played on yellow card occurrences. For instance, a positive coefficient signifies an increased likelihood of yellow cards being issued in matches played in Italy and Spain compared to a reference country, while a negative coefficient for sup_implied indicates that higher implied bookmaker odds for total goals are associated with a decreased likelihood of yellow cards. These findings underscore the intricate interplay between various contextual factors, including the footballing culture of different nations and bookmakers' expectations, in shaping referee decisions. Additionally, the coefficient for kick_off_timecatEvening suggests that matches held in the evening may lead to a higher likelihood of yellow cards, implying that the timing of matches can influence player behavior and, subsequently, the referee's disciplinary actions. Overall, the coefficients provide valuable guidance for understanding the nuanced dynamics of yellow card issuance in football and can inform strategies aimed at minimizing their impact on team performance.

The performance of both negative binomial models were evaluated using RMSE and MAE as evaluation metrics.

Table 11: Model Performance of Negative Binomial Regression

Model	RMSE	MAE
AIC	1.973	1.585
BIC	1.974	1.585

Interestingly, like the Poisson regression models, both negative binomial models displayed remarkably similar performance. The AIC-based model yielded an RMSE of approximately 1.973 and an MAE of about 1.585, while the BIC-based model produced a very similar RMSE of approximately 1.974 and an MAE of about 1.585.

The striking similarity in performance between the AIC-based and BIC-based negative binomial models is notable. This suggests that the choice of model selection criteria, whether AIC or BIC, does not significantly impact the predictive accuracy of the negative binomial regression models for yellow cards. The fundamental predictors, such as country, implied support, implied target

goals, attendance value, and kick-off time, seem to be robust in their influence on the total number of yellow cards in football matches, as they consistently contributed to both models' predictive accuracy.

Furthermore, the negative binomial models offer an advantage over the Poisson models by accounting for overdispersion, which is a common characteristic of count data in sports events like football matches. The negative binomial distribution allows for greater flexibility in modeling the variance of the data, capturing the inherent variability in yellow card occurrences more accurately.

5.2 Modelling Red Cards

In this chapter, the results of predictive modeling for red cards in football matches using Poisson regression and negative binomial regression are presented. Two models for each regression method are produced, each selected based on different information criteria: AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). The objective was to determine which model provided a better fit for predicting the total number of yellow cards, and the performance was evaluated using key metrics, including RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error).

5.2.1 Poisson Regression

In this section, the results of the Poisson regression models for predicting red cards in football matches is explored. Red cards are a critical event in football, and understanding the factors that influence their occurrence can have significant implications for both teams and players. Two Poisson regression models were employed, one based on AIC and another based on BIC, to identify key predictors and assess predictive accuracy.

For the Poisson regression models predicting red cards, several predictors were initially considered, including country, competition level, implied target goals, and kick-off time. The AIC-based model retained all of these predictors in the final model.

Table 12: Summary of Best Model Based on Lowest AIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	-1.563e+00	1.913e-01	-8.172	3.04e-16
countryFrance	6.728e-01	6.623e-02	10.158	$<$ 2e-16
countryGermany	1.864e-01	6.759e-02	2.758	0.005810
countryItaly	7.334e-01	7.143e-02	10.267	$<$ 2e-16
countrySpain	6.506e-01	6.251e-02	10.407	$<$ 2e-16

Coefficients	Estimate	Standard Error	z value	p-value
competition_ levelL2	6.360e-02	4.757e-02	1.337	0.181229
sup_implied	9.669e-03	2.894e-02	0.334	0.738331
tg_implied	-1.880e-01	5.921e-02	-3.175	0.001498
attendance_ value	-1.119e-06	1.711e-06	-0.654	0.513086
kick_off_timeca tEvening	-1.405e-01	3.985e-02	-3.526	0.000422
matches	8.260e-04	5.402e-04	1.529	0.126244
referee_group Experienced	1.225e-01	8.412e-02	1.456	0.145345
referee_group Novice	1.231e-01	1.455e-01	0.846	0.397700
referee_group Veteran	3.754e-02	5.439e-02	0.690	0.490061

The coefficients derived from both the AIC-based and BIC-based Poisson regression models provide valuable insights into the factors influencing the issuance of red cards in football matches. Notably, the coefficients for country indicators, such as France, Germany, Italy, and Spain, reveal significant variations in the likelihood of red card issuance based on the country where the match is played. For example, positive coefficients for France, Italy, and Spain indicate an increased probability of red cards being shown in matches held in these countries compared to a reference country. Conversely, a negative coefficient for the variable "tg_implied" suggests that higher implied bookmaker odds for total goals are associated with a decreased likelihood of red cards, indicating that matches with lower expected goal totals may be more prone to disciplinary actions. The coefficients for referee experience levels provide additional insights, with the "referee_group Experienced" coefficient suggesting a potential influence of referee experience on red card decisions, albeit with a relatively high p-value. Overall, these coefficients shed light on the complex interplay of factors, including geographical location, match dynamics, and referee experience, in shaping the issuance of red cards during football matches, contributing to a deeper understanding of this disciplinary aspect of the game.

While the BIC-based model selected a more parsimonious model, excluding competition level.

Table 13: Summary of Best Model Based on Lowest BIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	-1.28431	0.14104	-9.106	$<$ 2e-16
countryFrance	0.64686	0.06044	10.702	$<$ 2e-16
countryGermany	0.17725	0.06606	2.683	0.007288
countryItaly	0.69655	0.06093	11.433	$<$ 2e-16
countrySpain	0.62543	0.05754	10.870	$<$ 2e-16
tg_implied	-0.23490	0.05093	-4.612	3.99e-06
kick_off_timecatEvening	-0.13731	0.03971	-3.458	0.000544

The coefficients derived from the BIC-based Poisson regression model, which selected a more parsimonious representation by excluding competition level as a predictor, offer insights into the factors influencing the issuance of red cards in football matches. Notably, the coefficients for country indicators, such as France, Germany, Italy, and Spain, reveal significant variations in the likelihood of red card issuance based on the country where the match is held. Positive coefficients for France, Italy, and Spain suggest an increased probability of red cards being shown in matches played in these countries compared to the reference country. Conversely, the negative coefficient for the variable "tg_implied" implies that higher implied bookmaker odds for total goals are associated with a decreased likelihood of red cards. This finding suggests that matches with lower expected goal totals may be more prone to disciplinary actions. Additionally, the coefficient for "kick_off_timecat Evening" suggests that evening kick-off times are associated with a decreased likelihood of red cards compared to other times of the day. Overall, these coefficients provide valuable insights into the intricate interplay of geographical location, match dynamics, and implied goal expectations in influencing the issuance of red cards during football matches, even in a more parsimonious model without competition level as a predictor.

To evaluate the performance of both Poisson models, two common evaluation metrics were employed.

Table 14: Model Performance of Poisson Regression

Model	RMSE	MAE
AIC	0.459	0.340
BIC	0.459	0.340

Remarkably, both models demonstrated nearly identical performance, which suggests that the choice between AIC and BIC did not significantly affect the

models' predictive accuracy. The AIC-based model yielded an RMSE of approximately 0.459 and an MAE of about 0.340, while the BIC-based model produced very similar results with an RMSE of approximately 0.459 and an MAE of about 0.340.

The close alignment in predictive performance between the AIC-based and BIC-based Poisson models indicates that the core predictors, such as country, implied target goals, and kick-off time, are robust in explaining the occurrence of red cards in football matches. These findings are consistent with the notion that these variables capture essential aspects of match dynamics and player behavior that influence the likelihood of a red card.

The decision to include competition level in the AIC-based model suggests that this variable may have some influence on red card occurrences, but its exclusion in the BIC-based model highlights the trade-off between model complexity and parsimony. Researchers and practitioners can consider this trade-off when choosing a model for predicting red cards, depending on the specific research objectives and the importance of model interpretability.

Both AIC and BIC-based models provided nearly identical predictive accuracy. This suggests that researchers and sports analysts can confidently choose either model selection criterion, depending on their preferences for model complexity and interpretability.

5.2.2 Negative Binomial Regression

In this section, the results of negative binomial regression models for predicting red cards in football matches is looked into. Two negative binomial regression models were employed, one based on AIC and another based on BIC, to identify the key predictors and assess predictive accuracy.

Similar to previous models, model begin with considering several predictors, including country, competition level, implied target goals, and kick-off time. The AIC-based model retained all these predictors in the final model.

Table 15: Summary of Best Model Based on Lowest AIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	-1.42858	0.16329	-8.749	$<$ 2e-16
countryFrance	0.66395	0.06279	10.574	$<$ 2e-16
countryGermany	0.18091	0.06753	2.679	0.007390
countryItaly	0.72485	0.06449	11.240	$<$ 2e-16
countrySpain	0.63606	0.05937	10.714	$<$ 2e-16
competition_levelL2	0.07764	0.04022	1.930	0.053587

Coefficients	Estimate	Standard Error	z value	p-value
tg_implied	-0.19787	0.05577	-3.548	0.000389
kick_off_timecat Evening	-0.14127	0.04111	-3.436	0.000590

The coefficients derived from the AIC-based negative binomial regression model shed light on the factors influencing the occurrence of red cards during football matches. Notably, the coefficients for country indicators, including France, Germany, Italy, and Spain, indicate substantial variations in the likelihood of red card issuance based on the country where the match is hosted. Positive coefficients for France, Italy, and Spain suggest a higher probability of red cards being shown in matches held in these countries compared to the reference country. Conversely, the negative coefficients for "tg_implied" and "kick_off_timecat Evening" suggest that higher implied bookmaker odds for total goals and evening kick-off times are associated with a decreased likelihood of red cards. These findings offer valuable insights into the complex interplay of geographical location, match dynamics, and implied goal expectations in influencing the issuance of red cards in football matches, as identified by the AIC-based model.

While the BIC-based model selected a simpler version of the model, excluding competition level. This selection process allowed us to explore how model complexity affects predictive performance.

Table 16: Summary of Best Model Based on Lowest BIC

Coefficients	Estimate	Standard Error	z value	p-value
Intercept	-1.28494	0.14514	-8.853	$<$ 2e-16
countryFrance	0.64679	0.06214	10.409	$<$ 2e-16
countryGermany	0.17755	0.06752	2.630	0.008550
countryItaly	0.69625	0.06276	11.093	$<$ 2e-16
countrySpain	0.62567	0.05911	10.585	$<$ 2e-16
tg_implied	-0.23468	0.05243	-4.476	7.62e-06
kick_off_timecat Evening	-0.13724	0.04103	-3.344	0.000825

The coefficients derived from the BIC-based negative binomial regression model offer insights into the predictors that significantly influence the occurrence of red cards in football matches. Notably, the model's simplicity, achieved by excluding the "competition level" predictor, allows us to explore the impact of model

complexity on predictive performance. Positive coefficients for country indicators, including France, Germany, Italy, and Spain, indicate varying degrees of influence based on the host country. Specifically, matches held in France, Italy, and Spain are associated with a higher probability of red card issuance compared to the reference country. Conversely, the negative coefficients for "tg_implied" and "kick_off_timecat Evening" suggest that higher implied bookmaker odds for total goals and evening kick-off times are linked to a reduced likelihood of red cards being shown. These findings illuminate the nuanced interplay between geographical location, match dynamics, and implied goal expectations in shaping red card incidents during football matches, as discerned by the BIC-based model's streamlined approach.

To evaluate the performance of both Poisson models, two common evaluation metrics were employed.

Table 17: Model Performance of Negative Binomial Regression

Model	RMSE	MAE
AIC	0.459	0.340
BIC	0.459	0.340

Astonishingly, the AIC-based and BIC-based models produced nearly identical results, indicating that the choice of model selection criterion did not significantly influence the models' predictive accuracy. The AIC-based model yielded an RMSE of approximately 0.459 and an MAE of about 0.340, while the BIC-based model delivered similar results with an RMSE of approximately 0.459 and an MAE of about 0.340.

The consistency in predictive performance between the AIC-based and BIC-based negative binomial models suggests that the core predictors, such as country, implied target goals, and kick-off time, remain robust in explaining the occurrence of red cards in football matches. These findings reinforce the idea that these variables capture essential aspects of match dynamics and player behavior that significantly impact the likelihood of a red card.

The exclusion of competition level in the BIC-based model highlights the trade-off between model complexity and interpretability. While the AIC-based model suggests that competition level may have some influence on red card occurrences, the BIC-based model favors a simpler model without it. This choice should be guided by the research objectives and the need for a more parsimonious model.

5.3 Testing Models on Yellow and Red Cards During COVID-19

5.3.2 Yellow Cards

For predicting yellow cards during the COVID-19 period, both Poisson and negative binomial regression models that had been previously trained on data from before the pandemic were applied. Surprisingly, the RMSE and MAE metrics for both types of models remained remarkably consistent with their pre-COVID counterparts. This suggests that the factors influencing yellow card occurrences, such as country, implied target goals, and kick-off time, continued to exhibit similar patterns during the pandemic. The AIC and BIC-based models produced virtually identical results, indicating the robustness of these models across different time periods.

Table 18: Model Performance of Poisson Regression

Model	RMSE	MAE
AIC	1.968	1.583
BIC	1.967	1.583

Table 19: Model Performance of Negative Binomial Regression

Model	RMSE	MAE
AIC	1.968	1.583
BIC	1.967	1.583

5.3.3 Red Cards

Similarly, for red card prediction during the COVID-19 period, both Poisson and negative binomial regression models displayed consistent predictive performance compared to their pre-COVID results. The RMSE and MAE metrics once again remained stable, reinforcing the notion that the same predictors—country, implied target goals, and kick-off time—play a significant role in explaining red card occurrences, even in the face of unprecedented disruptions like a global pandemic. Both AIC and BIC-based models delivered similar levels of accuracy, indicating their resilience in adapting to changing circumstances.

Table 20: Model Performance of Poisson Regression

Model	RMSE	MAE
AIC	0.459	0.338
BIC	0.459	0.337

Table 21: Model Performance of Negative Binomial Regression

Model	RMSE	MAE
AIC	0.459	0.338
BIC	0.459	0.337

These findings provide valuable insights into the stability of the predictive models. The fact that they perform consistently well on data from different time periods suggests that the underlying factors affecting yellow and red card occurrences in football matches may be more enduring and less sensitive to external disruptions like COVID-19. This robustness allows teams and analysts to rely on these models for decision-making, even when faced with unexpected challenges that can impact the dynamics of the game.

CHAPTER SIX CONCLUSION

6.1 Summary and Conclusion

In this dissertation, we embarked on a data-driven journey to predict yellow and red card occurrences in football matches. Leveraging the power of statistical modeling, we delved into the factors that influence these crucial events during games. Our analysis encompassed two significant regression techniques: Poisson and negative binomial regression models. We also explored the influence of model selection criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), to refine our models further.

For yellow card predictions, our investigation revealed that certain key features, including the country of the match, implied target goals, and kick-off time, significantly influence the likelihood of yellow cards being issued. Our models, based on both AIC and BIC, consistently demonstrated their ability to provide accurate predictions across different datasets, including those recorded during the COVID-19 pandemic. This robustness implies that the underlying dynamics of yellow card occurrences remain relatively stable, regardless of external disruptions.

Similarly, in the realm of red card predictions, our analysis confirmed the importance of factors like country, implied target goals, and kick-off time. Again, our models, based on AIC and BIC, maintained their predictive accuracy when applied to data collected during the COVID-19 period. This resilience underscores the enduring influence of these predictors on red card events in football matches.

In conclusion, our research has yielded valuable insights into the predictability of yellow and red card occurrences in football matches. We have demonstrated the effectiveness of Poisson and negative binomial regression models in capturing the complexities of these events. Furthermore, the consistency in model performance across distinct time periods, including the unprecedented challenges posed by the COVID-19 pandemic, emphasizes the stability of the underlying factors governing card issuance in the sport.

These findings have practical implications for various stakeholders in football, from team managers and coaches to analysts and statisticians. By understanding the enduring influence of certain match characteristics, teams can make more informed decisions about player behavior, tactics, and game strategy. Moreover, these models can serve as valuable tools for referees, helping them manage matches more effectively by anticipating and responding to potential card incidents.

As football continues to evolve and adapt to changing circumstances, our predictive models offer a reliable foundation for decision-making in this dynamic sport. They empower stakeholders to make data-driven choices, enhance the quality of play, and ensure fair competition. In the future, this research can

be extended to encompass even more granular data and a broader range of factors, providing an even deeper understanding of the intricacies of football match dynamics.

6.2 Limitations and Suggestions

While our research has provided valuable insights into the prediction of yellow and red card occurrences in football matches, it is important to acknowledge several limitations that may impact the generalizability and scope of our findings.

Firstly, our models rely on historical match data, and the predictive power of these models may be influenced by changes in the game over time. Football is a sport that continually evolves, with rule changes, player strategies, and referee interpretations all potentially affecting the dynamics of card issuance. Our models may not capture these nuances fully.

Secondly, our analysis primarily considers match-level factors, such as the country of the match, implied target goals, and kick-off time. While these features provide valuable insights, they do not encompass all potential influencers of card events. Individual player behaviors, team dynamics, and specific referee decisions may play critical roles in card issuance but were not included in our models due to data limitations.

Thirdly, our models assume stationarity, meaning that the factors influencing card events remain constant over time. This assumption may not hold in situations of significant change, such as the impact of the COVID-19 pandemic. While our models demonstrated robustness during this period, external disruptions can introduce unpredictable elements that may not be accounted for in our analysis.

To address these limitations and further enhance our understanding of card occurrences in football matches, future research endeavors could consider including data on individual player behaviors and characteristics could provide a more comprehensive view of card issuance. Factors such as player aggression, discipline history, and playing position may significantly influence the likelihood of receiving cards.

Secondly, it might be a good idea to delve deeper into referee-specific data, considering the background and experience of the officials involved. Referee decisions are subjective and can vary greatly, so understanding the impact of different referees on card events could be valuable.

Thirdly, it is important to explore more advanced machine learning techniques, such as deep learning and ensemble methods, to capture complex, nonlinear relationships between predictors and card events. These methods may uncover hidden patterns and interactions that traditional regression models might miss.

Incorporating real-time data streams during matches could also enable more dynamic and accurate predictions. Tracking player movements, ball possession, and other in-game variables may offer richer insights into card occurrences.

Considering additional contextual factors, such as the importance of the match, rivalries, and fan behavioris an also an option. These elements can create unique dynamics that impact card issuance and warrant further investigation.

It is also possible to analyze the impact of external events, such as pandemics or political unrest, on card occurrences. Understanding how these disruptive factors influence the game can provide insights into football's resilience and adaptability.

Lastly, extending the validation of predictive models to different football leagues and competitions to assess the transferability of findings. Each league may have its own unique characteristics that affect card events.

By addressing these suggestions, future research can build upon our foundation and contribute to a more comprehensive and nuanced understanding of card occurrences in football matches, ultimately benefiting teams, referees, and fans alike.

Appendices

library(readxl) library(dplyr) library(tidyr) library(ggplot2) library(lubridate) library(knitr) library(kableExtra) library(randomForest) library(MASS)

dataset <- read_excel(params$data) dataset <- subset(dataset, select = -c(season, team1_name, team2_name))

glimpse(dataset)

anyDuplicated(dataset)

unique(dataset$country)$unique(dataset$referee)

colSums(is.na(dataset[sapply(dataset, is.numeric)]))

apply(dataset, 2, function(col) sum(col == "NA", na.rm = TRUE))

dataset$kick_off_datetime < -ymd_hms(dataset$kick_off_datetime)

dataset$kick_off_date < -as.Date(dataset$kick_off_datetime)

dataset <- subset(dataset, select = -c(kick_off_datetime))

dataNA <- read_excel("Data (variables to use) - NA.xlsx")

missing_by_country <- table(dataNA$country)

print(missing_by_country) plot(missing_by_country)

mode_referee_by_country <- dataset %>% group_by(country) %>% summarize(mode_referee = names(sort(table(referee), decreasing = TRUE))[1]) print(mode_referee_by_country)

dataset <- left_join(dataset, mode_referee_by_country, by = "country")

dataset$referee[dataset$referee == "NA"] <- dataset$mode_referee[dataset$referee == "NA"]

second_mode_referee <- dataset %>% filter(country == "Germany" & referee != "NA") %>% count(referee) %>% arrange(desc(n)) %>% slice(2) %>% pull(referee) print(second_mode_referee)

dataset$referee[dataset$country == "Germany" & dataset$referee == "NA"] <- second_mode_referee

dataset$mode_referee <- NULL

sum(is.na(dataset$sup_implied))$sum(is.na(dataset$tg_implied))

summary(dataset$sup_implied)$hist(dataset$sup_implied)

summary(dataset$tg_implied)$hist(dataset$tg_implied)

mean_sup_implied <- mean(dataset$sup_implied, na.rm = TRUE)$mean_tg_implied < -mean(dataset$tg_implied, na.rm = TRUE)

dataset$sup_implied[is.na(dataset$sup_implied)] <- mean_sup_implied dataset$tg_implied[is.na(dataset$tg_implied)] <- mean_tg_implied

dataset$total_yc < -dataset$team1_yc + dataset$team2_ycdataset$total_rc <- dataset$team1_rc + dataset$team2_rc

dataset <- subset(dataset, select = -c(team1_yc, team1_rc, team2_yc, team2_rc))

referee_summary <- dataset %>% group_by(referee) %>% summarise(matches = n()) %>% summarise(mean_matches = mean(matches), median_matches = median(matches), sd_matches = sd(matches), min_matches = min(matches), max_matches = max(matches))

referee_counts <- dataset %>% group_by(referee) %>% summarise(matches = n())

ggplot(referee_counts, aes(x = matches)) + geom_histogram(binwidth = 20, fill = "dodgerblue", color = "black") + labs(title = "Distribution of Matches per Referee", x = "Number of Matches", y = "Frequency")

quartiles <- quantile(referee_counts$matches, probs = c(0.25, 0.50, 0.75))

print(quartiles)

library(dplyr)

referee_counts <- referee_counts %>% mutate(referee_group = case_when(matches < 18 ~ "Novice", matches >= 18 & matches <= 54 ~ "Experienced", matches >= 55 & matches <= 120 ~ "Veteran", matches > 120 ~ "Elite", TRUE ~ NA_character_))

dataset <- left_join(dataset, referee_counts, by = "referee")

dataset <- subset(dataset, select = -c(referee))

dataset$kick_off_timecat < -ifelse(dataset$kick_off_timecat == 1, "Afternoon", "Evening")

dataset$competition_level < -ifelse(dataset$competition_level == 1, "L1", "L2")

dataset$kick_off_date < -as.Date(dataset$kick_off_date)

dataset$kick_off_year < -format(dataset$kick_off_date, "%Y") dataset$kick_off_month < -format(dataset$kick_off_date, "%m")

attendance_by_month_year <- dataset %>% group_by(kick_off_year, kick_off_month) %>% summarise(total_attendance = sum(attendance_value))

ggplot(attendance_by_month_year, aes(x = kick_off_month, y = total_attendance, fill = kick_off_year)) + geom_bar(stat = "identity", position = "dodge") + labs(title = "", x ="Month", y = "Total Attendance", fill = "Year") + theme_minimal()

dataset$kick_o ff_d ate$ < $-as.Date(dataset$kick_off_date) dataset$period$ < $-ifelse(dataset$kick_off_date <= as.Date("2020-03-31"), "Period 1", "Period 2")

period1_data <- dataset %>% filter(kick_off_date <= as.Date("2020-03-31"))

library(caret)

set.seed(123) # For reproducibility train_indices <- createDataPartition(period1_data$attendance_value, p = 0.7, list = FALSE) training_data <- period1_data[train_indices,] validation_data <- period1_data[-train_indices,]

period2_data <- dataset %>% filter(kick_off_date > as.Date("2020-03-31"))

dataset_yc <- subset(dataset, select = -c(total_rc,kick_off_date, kick_off_year,kick_off_month,period))
dataset_rc <- subset(dataset, select = -c(total_yc,kick_off_date, kick_off_year,kick_off_month,period))

training_data <- subset(training_data, select = -c(kick_off_date,kick_off_month,kick_off_year,period))
validation_data <- subset(validation_data, select = -c(kick_off_date,kick_off_month,kick_off_year,period))
period2_data <- subset(period2_data, select = -c(kick_off_date,kick_off_month,kick_off_year,period))

trainingyc_data <- subset(training_data, select = -c(total_rc)) trainingrc_data <- subset(training_data, select = -c(total_yc))

validationyc_data <- subset(validation_data, select = -c(total_rc)) validationrc_data <- subset(validation_data, select = -c(total_yc))

period2yc_data <- subset(period2_data, select = -c(total_rc)) period2rc_data <- subset(period2_data, select = -c(total_yc))

library(MASS)

ptyc <- glm(total_yc ~ .,data=trainingyc_data,family="poisson") stepwise_model_ptyc <- step(ptyc, direction = "both", trace = 0) summary(stepwise_model_ptyc) summary(ptyc) AIC(stepwise_model_ptyc) AIC(ptyc) stepwise_model_bic <- stepAIC(ptyc, direction = "both", trace = 0, k = log(nrow(trainingyc_data)))

summary(stepwise_model_bic) bic_selected_model <- BIC(stepwise_model_bic)

predictions_aic <- predict(stepwise_model_ptyc, newdata = validationyc_data, type = "response") predictions_bic <- predict(stepwise_model_bic, newdata = validationyc_data, type = "response")

rmse_aic <- sqrt(mean((validationyc_data$total_yc - predictions_a ic)^2)) $rmse_b ic$ < $-sqrt(mean((validationyc_d ata$total_yc - predictions_bic)^2))

mae_aic <- mean(abs(validationyc_data$total_yc - predictions_a ic)) $mae_b ic$ < $-mean(abs(validationyc_d ata$total_yc - predictions_bic))

print(paste("RMSE for AIC-based model:", rmse_aic)) print(paste("RMSE for BIC-based model:", rmse_bic))

print(paste(“MAE for AIC-based model:”, mae_aic)) print(paste(“MAE for BIC-based model:”, mae_bic))

nbtyc_model <- glm.nb(total_yc ~ ., data = trainingyc_data) summary(nbtyc_model)

nbtyc_model <- glm.nb(total_yc ~ .,data=trainingyc_data) stepwise_nbtyc_model <- step(nbtyc_model, direction = “both”, trace = 0) summary(stepwise_nbtyc_model) summary(nbtyc_model) AIC(stepwise_nbtyc_model) AIC(nbtyc_model)

stepwise_nbtyc_model_bic <- stepAIC(nbtyc_model, direction = “both”, trace = 0, k = log(nrow(trainingyc_data))) summary(stepwise_nbtyc_model_bic) bic_selected_nbtyc_model <- BIC(stepwise_nbtyc_model_bic)

predictions_aic_nbtyc_model <- predict(stepwise_nbtyc_model, newdata = validationyc_data, type = “response”) predictions_bic_nbtyc_model <- predict(stepwise_nbtyc_model_bic, newdata = validationyc_data, type = “response”)

rmse_nbtyc_model_aic <- sqrt(mean((validationyc_data$total_yc - predictions_aic)^2)) rmse_nbtyc_model_bic <- sqrt(mean((validationyc_data$total_yc - predictions_bic)^2)) mape_aic_nbtyc_model <- mean(abs((validationyc_data$total_yc - predictions_aic_nbtyc_model)/validationyc_data$total_yc)) * 100 mape_bic_nbtyc_model <- mean(abs((validationyc_data$total_yc - predictions_bic_nbtyc_model)/validationyc_data$total_yc)) * 100

mae_aic_nbtyc_model <- mean(abs(validationyc_data$total_yc - predictions_aic_nbtyc_model)) mae_bic_nbtyc_model <- mean(abs(validationyc_data$total_yc - predictions_bic_nbtyc_model))

print(paste(“RMSE for AIC-based model:”, rmse_nbtyc_model_aic)) print(paste(“RMSE for BIC-based model:”, rmse_nbtyc_model_bic)) print(paste(“MAPE for AIC-based model:”, mape_aic_nbtyc_model, “%”)) print(paste(“MAPE for p-value-based model:”, mape_bic_nbtyc_model, “%”))

print(paste(“MAE for AIC-based model:”, mae_aic_nbtyc_model)) print(paste(“MAE for p-value-based model:”, mae_bic_nbtyc_model))

ptrc <- glm(total_rc ~ .,data=trainingrc_data,family=“poisson”) stepwise_model_ptrc <- step(ptrc, direction = “both”, trace = 0) summary(stepwise_model_ptrc) summary(ptrc) AIC(stepwise_model_ptrc) AIC(ptrc) stepwise_model_bic <- stepAIC(ptrc, direction = “both”, trace = 0, k = log(nrow(trainingrc_data)))

summary(stepwise_model_bic) bic_selected_model <- BIC(stepwise_model_bic)

predictions_aic <- predict(stepwise_model_ptrc, newdata = validationrc_data, type = “response”) predictions_bic <- predict(stepwise_model_bic, newdata = validationrc_data, type = “response”)

rmse_aic <- sqrt(mean((validationrc_data$total_rc - predictions_aic)^2)) rmse_bic <- sqrt(mean((validationrc_data$total_rc - predictions_bic)^2))

mae_aic <- mean(abs(validationrc_data$total_rc - predictions_aic))mae_bic <
−mean(abs(validationrc_data$total_rc - predictions_bic))

print(paste("RMSE for AIC-based model:", rmse_aic)) print(paste("RMSE for BIC-based model:", rmse_bic)) print(paste("MAPE for AIC-based model:", mape_aic, "%")) print(paste("MAPE for p-value-based model:", mape_bic, "%"))

print(paste("MAE for AIC-based model:", mae_aic)) print(paste("MAE for p-value-based model:", mae_bic))

nbtrc_model <- glm.nb(total_rc ~ ., data = trainingrc_data) summary(nbtrc_model)

stepwise_nbtrc_model <- step(nbtrc_model, direction = "both", trace = 0) summary(stepwise_nbtrc_model) summary(nbtrc_model) AIC(stepwise_nbtyc_model) AIC(nbtyc_model)

stepwise_nbtrc_model_bic <- stepAIC(nbtrc_model, direction = "both", trace = 0, k = log(nrow(trainingrc_data))) summary(stepwise_nbtrc_model_bic) bic_selected_nbtyc_model <- BIC(stepwise_nbtyc_model_bic)

predictions_aic_nbtrc_model <- predict(stepwise_nbtrc_model, newdata = validationrc_data, type = "response") predictions_bic_nbtrc_model <- predict(stepwise_nbtrc_model_bic, newdata = validationrc_data, type = "response")

rmse_nbtrc_model_aic <- sqrt(mean((validationrc_data$total_rc−predictions_aic)^2))rmse_nbtrc_model_bic <
−sqrt(mean((validationrc_data$total_rc - predictions_bic)^2))

mae_aic_nbtrc_model <- mean(abs(validationrc_data$total_rc−predictions_aic))mae_bic_nbtrc_model <
−mean(abs(validationrc_data$total_rc - predictions_bic))

print(paste("RMSE for AIC-based model:", rmse_nbtrc_model_aic)) print(paste("RMSE for BIC-based model:", rmse_nbtrc_model_bic)) print(paste("MAPE for AIC-based model:", mape_aic_nbtrc_model, "%")) print(paste("MAPE for p-value-based model:", mape_bic_nbtrc_model, "%"))

print(paste("MAE for AIC-based model:", mae_aic_nbtrc_model)) print(paste("MAE for p-value-based model:", mae_bic_nbtrc_model))

predictions_aic <- predict(stepwise_model_ptyc, newdata = period2yc_data, type = "response") predictions_bic <- predict(stepwise_model_bic, newdata = period2yc_data, type = "response")

rmse_aic <- sqrt(mean((period2yc_data$total_yc − predictions_aic)^2))rmse_bic <
−sqrt(mean((period2yc_data$total_yc - predictions_bic)^2))

mae_aic <- mean(abs(period2yc_data$total_yc − predictions_aic))mae_bic <
−mean(abs(period2yc_data$total_yc - predictions_bic))

print(paste("RMSE for AIC-based model:", rmse_aic)) print(paste("RMSE for BIC-based model:", rmse_bic))

print(paste(“MAE for AIC-based model:”, mae_aic)) print(paste(“MAE for BIC-based model:”, mae_bic))

predictions_aic_nbtyc_model <- predict(stepwise_nbtyc_model, newdata = period2yc_data, type = “response”) predictions_bic_nbtyc_model <- predict(stepwise_nbtyc_model_bic, newdata = period2yc_data, type = “response”)

rmse_nbtyc_model_aic <- sqrt(mean((period2yc_data$total_yc - predictions_aic)^2))$rmse_nbtyc_model_bic <- sqrt(mean((period2yc_data$total_yc - predictions_bic)^2))$

mae_aic_nbtyc_model <- mean(abs(period2yc_data$total_yc - predictions_aic_nbtyc_model))$mae_bic_nbtyc_model <- mean(abs(period2yc_data$total_yc - predictions_bic_nbtyc_model))$

print(paste(“RMSE for AIC-based model:”, rmse_nbtyc_model_aic)) print(paste(“RMSE for BIC-based model:”, rmse_nbtyc_model_bic))

print(paste(“MAE for AIC-based model:”, mae_aic_nbtyc_model)) print(paste(“MAE for BIC-based model:”, mae_bic_nbtyc_model))

predictions_aic <- predict(stepwise_model_ptrc, newdata = period2rc_data, type = “response”) predictions_bic <- predict(stepwise_model_bic, newdata = period2rc_data, type = “response”)

rmse_aic <- sqrt(mean((period2rc_data$total_rc - predictions_aic)^2))$rmse_bic <- sqrt(mean((period2rc_data$total_rc - predictions_bic)^2))$

mae_aic <- mean(abs(period2rc_data$total_rc - predictions_aic))$mae_bic <- mean(abs(period2rc_data$total_rc - predictions_bic))$

print(paste(“RMSE for AIC-based model:”, rmse_aic)) print(paste(“RMSE for BIC-based model:”, rmse_bic))

print(paste(“MAE for AIC-based model:”, mae_aic)) print(paste(“MAE for p-value-based model:”, mae_bic))

predictions_aic_nbtrc_model <- predict(stepwise_nbtrc_model, newdata = period2rc_data, type = “response”) predictions_bic_nbtrc_model <- predict(stepwise_nbtrc_model_bic, newdata = period2rc_data, type = “response”)

rmse_nbtrc_model_aic <- sqrt(mean((period2rc_data$total_rc - predictions_aic)^2))$rmse_nbtrc_model_bic <- sqrt(mean((period2rc_data$total_rc - predictions_bic)^2))$

mae_aic_nbtrc_model <- mean(abs(period2rc_data$total_rc - predictions_aic))$mae_bic_nbtrc_model <- mean(abs(period2rc_data$total_rc - predictions_bic))$

print(paste(“RMSE for AIC-based model:”, rmse_nbtrc_model_aic)) print(paste(“RMSE for BIC-based model:”, rmse_nbtrc_model_bic)) print(paste(“MAE for AIC-based model:”, mae_aic_nbtrc_model)) print(paste(“MAE for p-value-based model:”, mae_bic_nbtrc_model))

References

Agresti, A. (2006). An introduction to categorical data analysis: Second edition. In *An Introduction to Categorical Data Analysis: Second Edition*. https://doi.org/10.1002/0470114754

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, *37*. https://doi.org/10.1080/02664760802684177

Borland, J. F., & MacDonald, B. (2020). Football without fans? Sport in a time of pandemic. *Managing Sport and Leisure*, 1–7.

Clarke, S. R., & Norman, J. M. (2019). Data analytics and sport: Increasing performance, enhancing strategy, and improving decision making. *Journal of Sports Sciences*, *37*, 1578–1583.

Decroos, T., Haaren, J. V., Davis, J., & Deprez, K. (2019). Actions speak louder than goals: Valuing player actions in soccer. *Journal of Sports Analytics*, *5*, 185–193.

Divos, P. (2020). *Modelling of the in-play football betting market*. UCL (University College London).

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *46*. https://doi.org/10.1111/1467-9876.00065

Fernandez-Corugedo, E., & McMahon, M. (2021). *COVID-19 and the market for football players* (p. =).

Isaiah. (2023). *Asian handicap explained*. https://www.betshoot.com/betting-guides/asian-handicap-betting/

Islam, M. A., Kabir, M. R., & Talukder, A. (2020). Triggering factors associated with the utilization of antenatal care visits in bangladesh: An application of negative binomial regression model. *Clinical Epidemiology and Global Health*, *8*, 1297–1301. https://doi.org/10.1016/j.cegh.2020.04.030

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*, 381–393.

Kovalchik, S., Schulte, E., & Unkelbach, C. (2016). Data-driven identification of risk factors for yellow card accumulation in soccer. *Journal of Sports Sciences*, *34*, 1340–1347.

Mohan, N., & Samuelsen, B. M. (2017). Modelling the number of goals scored by a team in a match of association football. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*, 831–850.

Peña, J., Touchette, H., & Hartley, C. (2012). The probability of receiving a red card: Do referees truly discriminate? In *The Beautiful Game? Searching for the Soul of Football* (pp. 163–174).