Chapter 1

Introduction

This project aims to delve into the realm of La Liga through a statistical lens, exploring various aspects such as team performance metrics and predictive modeling of match outcomes. The dataset comprises detailed statistical records of matches played in the Spanish La Liga over the past decade.

Label	**Description**
Date	Date of the match
HomeTeam	Home Team of the match
AwayTeam	Away Team of the match
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
FTR	Full Time Result (H = Home Win, D = Draw, A = Away Win)
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result (H = Home Win, D = Draw, A = Away Win)
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HC	Home Team Corners
AC	Away Team Corners
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards

Table 1.1: Dataset Acronyms

Chapter 2

Steps Involved in the Process

2.1 Data Acquisition

The dataset is sourced from `http://www.football-data.co.uk/`, providing comprehensive match information.

2.2 Data Preparation and Cleaning

The dataset spans from the 2009/2010 season to 2022/2023 season. Key cleaning steps include:

Data Cleaning Steps

- **Missing Values:** Checked for and handled missing values.
- **Data Types:** Converted necessary columns to appropriate data types.

2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is crucial for understanding the dataset. It includes visualizing trends, identifying patterns, and summarizing statistics.

Analysis

Univariate analysis included a bar plot visualizing the distribution of match outcomes (Home Win, Draw, Away Win) (see Figure 2.1). Bivariate analysis involved boxplots to examine the relationships between goals scored and shots taken by home teams with match outcomes (see Figure 2.2).

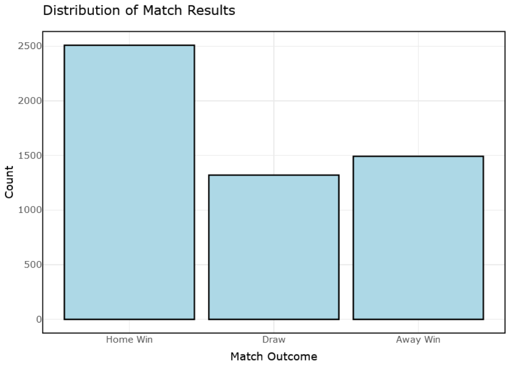

Figure 2.1: Match Results Visualization

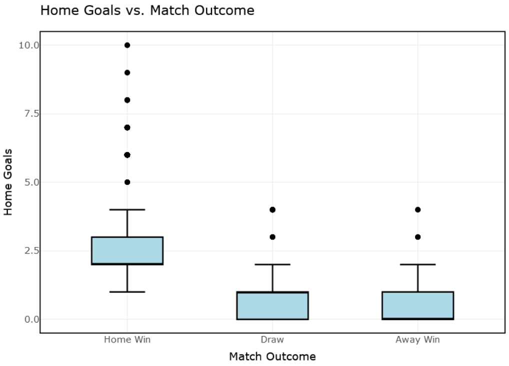

Figure 2.2: Home Team Goals vs. Match Outcome

2.4 Remove Multicollinear Variables

A heatmap visualized the correlation strengths, with green indicating positive, red negative, and white minimal correlation. Highly correlated variables were identified and removed, recalculating the correlation matrix to reduce multicollinearity and enhance the reliability of regression models by eliminating variables that could cause instability or inflate standard errors.

2.5 Model Building

This phase involves developing statistical models to analyze and predict match outcomes.

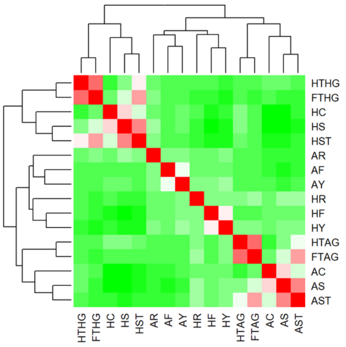

Figure 2.3: Correlation Matrix with Multicollinear Variables

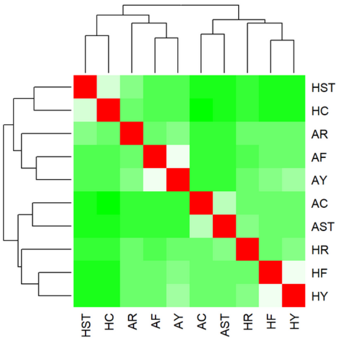

Figure 2.4: Removed Multicollinear Variables

2.5.1 Logistic Regression

Logistic regression was used to predict match outcomes using home and away team statistics and match features, with model fit assessed via the Akaike Information Criterion (AIC), where lower values indicate better models. The Variance Inflation Factor (VIF) was calculated to evaluate multicollinearity, revealing potential issues with high VIF values. Interaction terms (HTHG × HTAG, HST × AST, HR × AR, and HY × AY) were subsequently added to the model. The refitted model indicated that HST x AST and HY x AY were not significant, leading to their removal. AIC values were computed for four models: the full model, the model with removed multicollinearity, the model with added interaction terms to the full model, and the model with added interaction terms to the cleaned model, confirming that the initial full model remained the best.

Model	**df**	**AIC**
`multinom_model`	34	5286.169
`multinom_model.inter`	38	5288.942
`multinom_model.clean.inter`	30	5309.557
`multinom_model.clean`	22	5316.781

Table 2.1: AIC and Degrees of Freedom for Various Multinomial Models

2.6 Model Comparison

Let's compare all the models done so far in terms of prediction power to be able to compare them with other types of models:

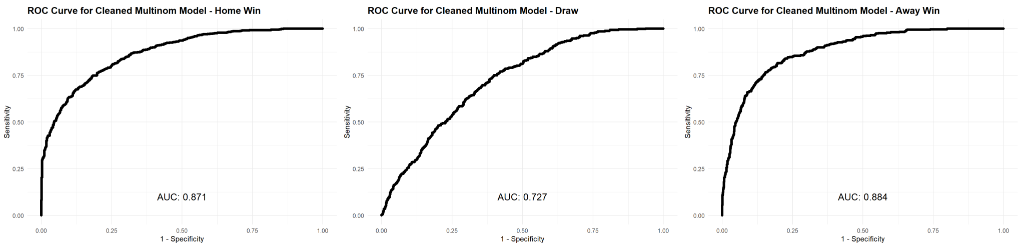

Figure 2.5: ROC Curves for Cleaned Multinomial Model

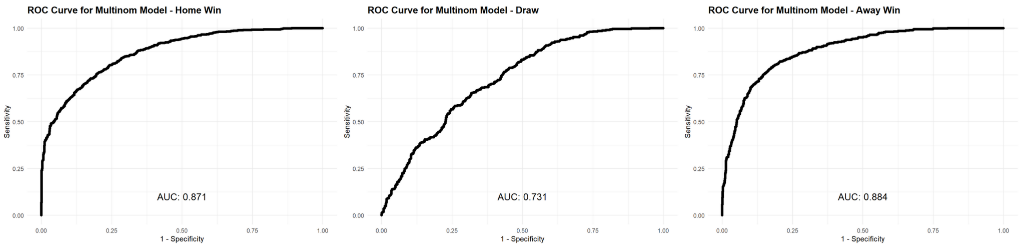

Figure 2.6: ROC Curves for Multinomial Model

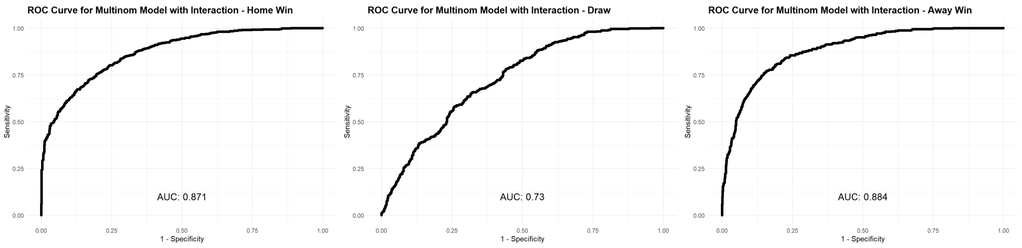

Figure 2.7: ROC Curves for Multinomial Model with Interaction Terms

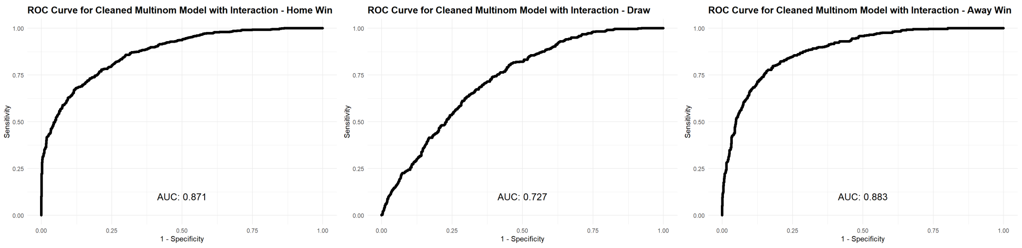

Figure 2.8: ROC Curves for Cleaned Multinomial Model with Interaction Terms

2.6.1 Stepwise Model

The stepwise model selection was performed using the ‘stepAIC‘ function, which iteratively adds and removes predictors based on the Akaike Information Criterion (AIC) to optimize the model. The process was conducted in both forward and backward directions, with progress displayed at each step. The final model resulted in an AIC of 2339.505, which was better than all the other models.

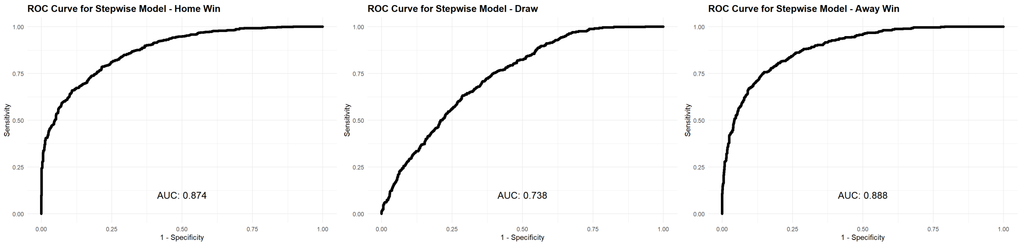

Figure 2.9: ROC Curves for Stepwise Model

	Away Win	**Draw**	**Home Win**
Away Win	311	91	48
Draw	71	132	82
Home Win	65	167	622

Table 2.2: Stepwise Model Confusion Matrix

2.6.2 Lasso Regression

The data was prepared by creating a design matrix from the training dataset and scaling the predictors before fitting a multinomial logistic regression model using LASSO regularization with the glmnet package. Cross-validation was conducted with the ‘cv.glmnet‘ function to determine the optimal regularization parameter (λ), visualized through plots of coefficient paths and cross-validated errors. The optimal λ minimizing the mean square error was 0.001377978, while a more regularized value, λ_{1se}, was 0.01066918.

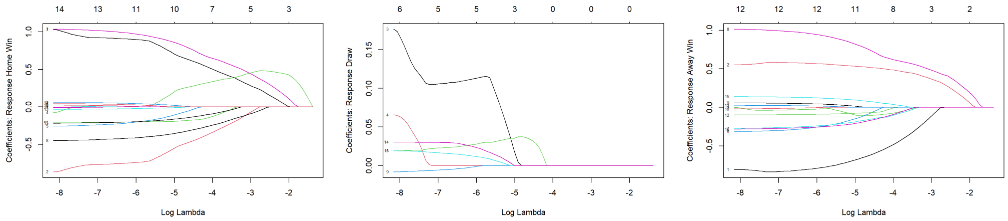

Figure 2.10: LASSO Plots

2.6.3 Ridge Logistic Regression

The ridge regression model was applied using all predictors. A multinomial ridge regression model was fitted using the ‘glmnet‘ function with ‘alpha = 0‘, and the coefficients were plotted against the regularization parameter (λ). Cross-validation via the ‘cv.glmnet‘ function determined the optimal λ, with the plot illustrating the cross-validated mean squared error for various λ values. The optimal value minimizing the error was $\lambda_{\text{min}} = 0.02522716$, while the selected $\lambda_{\text{1se}} = 0.04408518$ represents the simplest model within one standard error of the minimum. Predictions were then applied to the test set, resulting in the confusion matrix.

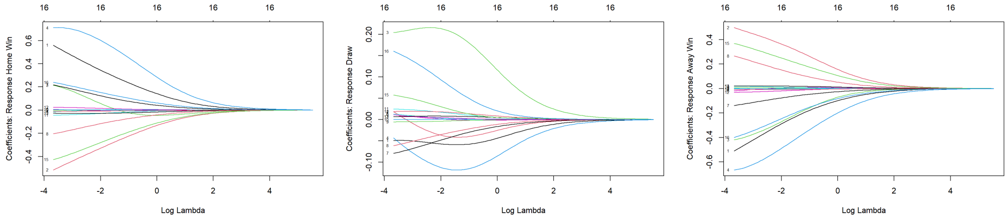

Figure 2.11: Ridge Plots

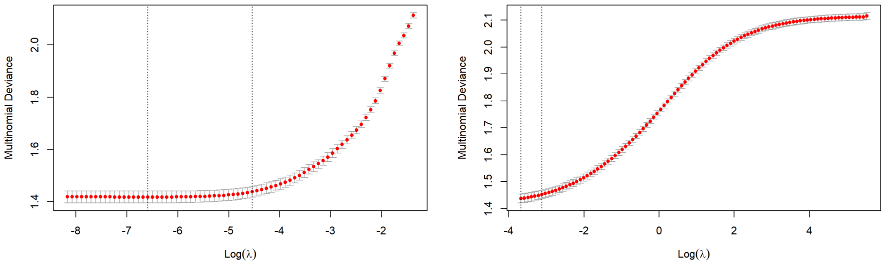

Figure 2.12: Best Log values for Lasso (left) and Ridge (right)

2.6.4 Linear Discriminant Analysis

This section applies Linear Discriminant Analysis (LDA), a generative model, using the full set of predictors. The LDA model is trained on the scaled training data to predict the target variable ‘FTR‘. Predictions are made on the test data, and a confusion matrix is generated to evaluate the model’s performance by comparing predicted classes with actual FTR values.

	Away Win	**Draw**	**Home Win**
Away Win	284	77	51
Draw	113	187	124
Home Win	50	132	577

Table 2.3: LDA Confusion Matrix

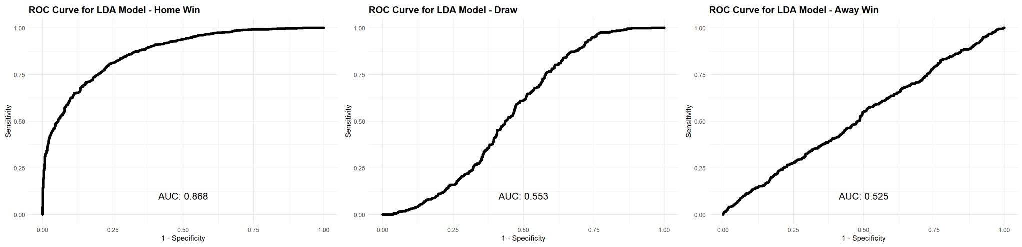

Figure 2.13: LDA ROC Curves

2.6.5 Naive Bayes

The Naive Bayes algorithm was employed to predict the Full-Time Result (FTR) based on the training data, assuming conditional independence of features given the class label. The model was then applied to the test dataset to predict class labels for FTR, and a confusion matrix was generated to compare predicted and actual classes.

	Away Win	**Draw**	**Home Win**
Away Win	286	100	66
Draw	115	184	142
Home Win	46	112	544

Table 2.4: Naive Bayes Confusion Matrix

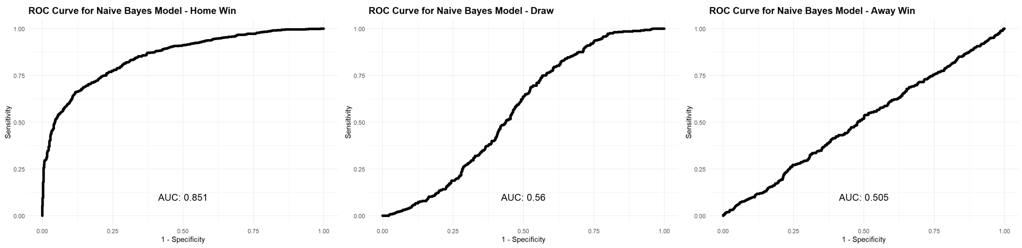

Figure 2.14: Naive Bayes ROC Curves

2.7 Conclusion

The best models in terms of different evaluation metrics for each class are summarized as follows:

- **Best Model for Accuracy:**
  - ”Away Win”: Cleaned Model
  - ”Draw”: Cleaned Interaction Model
  - ”Home Win”: Full Model
- **Best Model for Sensitivity:**
  - ”Away Win”: Cleaned Interaction Model
  - ”Draw”: Full Model
  - ”Home Win”: Ridge Model
- **Best Model for Specificity:**
  - ”Away Win”: LDA Model
  - ”Draw”: Ridge Model
  - ”Home Win”: Naive Bayes

One model that performs well across all conditions (Away Win, Draw, and Home Win) and balances accuracy, sensitivity, and specificity is the Cleaned Model. It provides an accuracy of 74.34% for Away Win, 73.56% for Draw, and 72.99% for Home Win. In terms of specificity, it achieves 74.13% for Away Win, 71.81% for

Model	**Class**	**Accuracy**	**Sensitivity**	**Specificity**
Full Model	Away Win	0.7299896	0.7885189	0.7148773
Full Model	Draw	0.7289446	0.8058413	0.6985173
Full Model	Home Win	0.7433647	0.7705527	0.7412823
Full Interaction Model	Away Win	0.7356322	0.7924588	0.7180400
Full Interaction Model	Draw	0.7299896	0.7885189	0.7148773
Full Interaction Model	Home Win	0.7289446	0.8058413	0.6985173
Cleaned Model	Away Win	0.7433647	0.7705527	0.7412823
Cleaned Model	Draw	0.7356322	0.7924588	0.7180400
Cleaned Model	Home Win	0.7299896	0.7885189	0.7148773
Cleaned Interaction Model	Away Win	0.7289446	0.8058413	0.6985173
Cleaned Interaction Model	Draw	0.7433647	0.7705527	0.7412823
Cleaned Interaction Model	Home Win	0.7356322	0.7924588	0.7180400
Stepwise Model	Away Win	0.6677116	0.6957494	0.8736934
Stepwise Model	Draw	0.6677116	0.3333333	0.8723937
Stepwise Model	Home Win	0.6677116	0.8271277	0.7247924
LASSO Model	Away Win	0.6683386	0.6890380	0.8780488
LASSO Model	Draw	0.6683386	0.3232323	0.8782319
LASSO Model	Home Win	0.6683386	0.8377660	0.7117438
Ridge Model	Away Win	0.6733542	0.6912752	0.8824042
Ridge Model	Draw	0.6733542	0.3257576	0.8807339
Ridge Model	Home Win	0.6733542	0.8457447	0.7117438
LDA Model	Away Win	0.6570533	0.6353468	0.8885017
LDA Model	Draw	0.6570533	0.4722222	0.8023353
LDA Model	Home Win	0.6570533	0.7672872	0.7841044
Naive Bayes	Away Win	0.6357367	0.6398210	0.8554007
Naive Bayes	Draw	0.6357367	0.4646465	0.7856547
Naive Bayes	Home Win	0.6357367	0.7234043	0.8125741

Table 2.5: Performance of Different Models by Class

Draw, and 71.43% for Home Win. For sensitivity, the Cleaned Model performs with 77.05% for Away Win, 79.25% for Draw, and 78.85% for Home Win.

While no single model dominates all metrics in each case, the Cleaned Model offers a strong balance across all three classes, making it a reliable choice. Therefore, the Cleaned Model is selected for predicting the final results of the 2023/2024 season matches.

The prediction was made using the data from the 2023/2024 season, which can be found in the folder `future_predictions/predictions`.