CREATING A BETTING STRATEGY

AKOS PLAJNER SIMON

MASTER THESIS
DATA SCIENCE & SOCIETY

STUDENT NUMBER

2111098

COMMITTEE

dr. Bruno Nicenboim

LOCATION

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science &

Artificial Intelligence

Tilburg, The Netherlands

DATE

May 20th, 2024

WORD COUNT

8045

CREATING A BETTING STRATEGY

AKOS PLAJNER SIMON

ABSTRACT

Predicting outcomes in football matches has long been a subject of interest, particularly within betting markets. While much attention has been devoted to forecasting match results, this study focuses on predicting the number of goals scored, specifically the popular over/under 2.5 goals market. Leveraging Machine Learning techniques on a dataset comprising 5236 matches, nine distinct betting strategies were developed and analyzed. The primary objective was to identify the strategy yielding the highest Return on Investment (ROI) and ascertain the key variables driving its success. Upon comparing various models, the combination of an Elastic-Net and a Random Forest model emerged with the highest ROI of 0.62%. Subsequently, an examination of the most influential variables revealed that factors such as odds, goals scored, and points significantly impacted the predictive accuracy of the selected models. Notably, limitations were acknowledged, including the absence of football-specific variables and the restricted scope of predictions to two rounds. Furthermore, an analysis of end-season games shed light on their impact on ROI, revealing a decline in performance when considering matches with lower stakes, where final league positions may already be determined, potentially affecting team motivation. Including these games in the dataset led to a further reduction in ROI. In addressing the research sub-questions, the study follows a logical progression, presenting betting strategy results, evaluating the top-performing models, and concluding with insights into the strategy for end-season games.

ACKNOWLEDGMENTS

I want to express my deepest gratitude to Dr. Bruno Nicenboim for his guidance, support, and scholarly insights throughout the journey of completing this thesis. Dr. Nicenboim's mentorship has been instrumental in shaping my research direction and refining my analytical skills. I am thankful for his time and effort to provide constructive feedback. Furthermore, I am grateful for the opportunities Dr. Nicenboim has provided for professional growth through his research seminars. His generosity in sharing his knowledge and expertise has enriched my academic experience and broadened my understanding of the field.

Lastly, I want to express my heartfelt gratitude to my friends and family. My academic journey would be unimaginable without their unwavering encouragement, sacrifices, and love.

ETHICS STATEMENT

This thesis did not involve collecting data from human participants or animals. The dataset utilized in this study was openly accessible to the public, and its ownership remains in the original domain both during and after the completion of this thesis.

All data used were obtained from publicly available sources and rigorously analyzed according to established academic standards. Care was taken to ensure the confidentiality and integrity of the dataset, and no personally identifiable information was included in the analysis.

Any modifications or transformations applied to the dataset were documented, and other researchers thoroughly described the methodology employed in the analysis to facilitate replication.

In adherence to principles of academic integrity, proper attribution was given to all sources consulted, including the dataset itself. Any limitations or biases inherent in the dataset were acknowledged and addressed to the best of my ability within the scope of this study.

This research aims to adhere to ethical guidelines and standards, prioritizing integrity, transparency, and respect for data ownership throughout the process.

1 INTRODUCTION

1.1 *Motivation*

This study arises from a profound interest in football and sports, particularly the challenge of predicting outcomes in the dynamic context of the football betting market. Building on prior research conducted during my BSc studies, where I investigated the creation of a successful betting strategy using econometric tools, this thesis aims to overcome the limitations of previous work.

While extensive research exists on predicting traditional outcomes like Home Win/Draw/Away Win (Choi et al., 2023; Constantinou, 2019; Heijboer, 2022), there is a noticeable gap in studies focusing on the total number of goals scored. By delving into this area and explicitly targeting the over/under 2.5 goals market, this research seeks to expand the range of options available to bettors.

1.2 *Societal and Scientific Contribution*

From a societal perspective, this research broadens the betting market's horizons by redirecting attention from game outcomes to the total number of goals. Such a shift could benefit regulators, bettors, and betting companies by providing additional insights for decision-making. Moreover, facilitating comparisons of strategies among participants could enrich understanding of betting behavior and market efficiency.

Scientifically, this research provides evidence regarding the effectiveness of machine learning models in formulating betting strategies. It also illuminates which machine learning algorithms are most effective in this context. Furthermore, it advances the exploration of risk management optimization in uncertain environments from a decision-making standpoint.

1.3 *About Betting Strategies in General*

Betting strategies do not have an exact definition, but they generally involve developing probabilities for various outcomes and systematically placing bets based on them. These strategies are applied iteratively, allowing their effectiveness to be evaluated after a series of trials.

A pivotal metric for assessing the success of a betting strategy is the return on investment (ROI). Calculating ROI, which is the ratio of the bettor's net gain and the total amount of bet (Ethier & Ethier, 2010), is considered the best predictor for evaluating the performance of a betting strategy. This method offers a precise and measurable evaluation of a strategy's performance over time.

It's important to note that there are numerous approaches to developing betting strategies, with probability calculation being a standard method (Angelini & De Angelis, 2017; Constantinou, 2019; Wheatcroft, 2020).

The ROI metric is a widely used and sophisticated tool for appraising strategy success, offering valuable insights into the effectiveness of different approaches (Boshnakov et al., 2017; Constantinou, 2019; Stübinger et al., 2019).

2 LITERATURE REVIEW

The goal of a strategy is to yield a profit; thus, the best indicator of its success is the return on investment (ROI). ROI can be calculated by determining the bettor's capital after several trials. The primary assumption is that the bettor starts with an initial capital and bets a fixed amount on each trial. Relevant formulas for calculating the bettor's capital after n trials and the ROI of a betting strategy are broadly applicable for evaluating performance (Ethier & Ethier, 2010).

The milestone paper of Constantinou (2019), the Dolores Model, performed exceptionally well in the English Premier League, producing an ROI of 20% over seven seasons, while the overall ROI (worldwide) was 1.09%. Constantinou built a Hybrid Bayesian Network that combined dynamic ELO ratings of football teams. ELO ratings, as shown by Glickman and Jones (1999), are primarily used in chess and provide a numeric rating designed to indicate the strength of a player based on their previous performance. Constantinou's model predicted football outcomes from 52 leagues worldwide for seven consecutive seasons. These results can be considered impressive, given the diversity of the leagues regarding division and the unpredictable nature of football.

Some approaches to predicting football outcomes involve logistic regression models, which estimate the probability of a particular event, such as the outcome of a match. Wheatcroft (2021) used two types of ratings: general attacking performance (GAP) and bivariate attacking performance (BAP). He hypothesized that predicting match statistics unknown before the game (such as the number of shots) would be the best way to predict outcomes. Therefore, he first predicted these match statistics and then used them to make predictions for betting items. This resulted in a two-layered prediction system that produced a profit of 4.88%.

The Hybrid Bayesian Network has delivered impressive results; however, some studies have explored more advanced machine learning techniques. For example, Rodrigues and Pinto (2022) used the English Premier League dataset covering four seasons to compare eight algorithms, including Random Forest (RF), Support Vector Machines (SVM), k-nearest Neighbors (KNN), and Logistic Regression. Initially, they used only 18 variables, and the SVM model emerged as the frontrunner, achieving an accuracy of 61.32% and an ROI of 24.4%. When interactions were incorporated, the Random Forest model proved to be the best, with an accuracy of 65.26% and an ROI of 53%. This demonstrates that, despite the notable performance of the Hybrid Bayesian Network, other machine learning algorithms can yield higher ROIs.

A similar approach was conducted by Mandadapu (2024), who experimented with data split over time and analyzed the performance of RF, SVM, KNN, and Gradient Boosting methods. Among these, SVM outperformed the other models across multiple performance metrics. Additionally, he concluded that feature selection is essential in model creation, and selecting the appropriate ratio of historical to recent data is crucial. Other studies have also widely used RF and SVM, demonstrating their effectiveness in football prediction (Baboota & Kaur, 2019). However, these papers are limited by the lack of information on ROIs or profits, meaning conclusions are based solely on performance metrics. He suggests predicting the number of goals scored and conceded to improve prediction accuracy and explore diverse betting systems.

Another approach involves using regression models to predict the season ranking of a team, highlighting the power of ElasticNet models in sports performance analysis (Rebbouj & Lotfi, n.d.). Although this study did not aim to create a betting strategy based on season rankings, it suggests that regression analysis is the most accurate method for predicting numerical scores in football. However, a limitation of this paper is that while predicting football season rankings is relevant, it may not be an effective method for predicting individual match events.

Many approaches to creating betting strategies focus on forecasting game outcomes, often neglecting other opportunities like predicting the total number of goals. While these methods can be profitable, they may limit market opportunities by overlooking specific events, such as betting on over/under 2.5 goals. Although some attention has been given to this area in scientific literature, it has not received as much focus as outcome prediction.

Successfully forecasting over and under 2.5 goals has been achieved using the previously mentioned general attacking performance (GAP) ratings (Wheatcroft, 2020). This method entails assigning GAP scores, constructing a logistic regression model, and calculating probabilities based on these scores. However, despite its historical profitability, recent years have seen a troubling trend: certain strategies have ceased yielding profits. The reasons behind this decline remain elusive, prompting a call for future research to investigate possible explanations for this phenomenon.

Not all approaches rely on classical machine learning tools; models such as the Weibull and Poisson have also been employed to forecast the total number of goals and construct betting strategies (Boshnakov et al., 2017).

Angelini and De Angelis (2017) introduced the PARX model, where they separately predict the number of home and away goals using the Poisson distribution to determine probabilities. Incorporating the PARX model's approach, which involves developing a Poisson autoregressive model with exogenous covariates and separating Home and Away Goals, are two essential attributes I intend to utilize in my work. The model generates predictions for home and away goals separately. Then, these outcomes are used to generate Poisson distributions, where the lambda value of the distribution is exactly the model output. For example, if the model output for the home team is 1.32, it becomes the lambda value of the Poisson distribution, thus indicating the probability of scoring:

- 0 goals = 26.7
- 1 goal = 35.3
- 2 goals = 23.3
- ...

Following this step, determining the probabilities of each match result becomes straightforward.

As evident, all these approaches apply to my work. Despite their suitability, further exploration is needed to predict the total number of goals using machine learning techniques. Moreover, my aspirations align with the literature concerning future work in this area.

3 RESEARCH STRATEGY AND RESEARCH QUESTIONS

In this paper, I will predict the well-known betting item of under and over 2.5 goals. Scoring 0, 1, or 2 goals in a football game is considered under 2.5, while scoring three or more goals is considered over 2.5. I will create betting strategies based on supervised machine-learning techniques and compare them based on profitability. This leads to my main research question.

When explaining the procedure for addressing the sub-questions, I will allocate more time to elucidating the first sub-question. This decision stems from the complexity of the subject matter and the need for some prior knowledge, which necessitates a more thorough explanation. By providing additional context and background information, I aim to ensure clarity and facilitate a better understanding of the subsequent inquiries. Consequently, I will provide a more concise explanation for the last two sub-questions, as they are inherently linked to the understanding established by elaborating on the first sub-question.

To what extent can a successful betting strategy be created using Machine Learning Models?

I must model the number of home and away goals to create betting strategies. After generating model predictions, I will calculate the probabilities of scoring over and under 2.5 goals in a game. Based on the work of Baboota and Kaur (2019), Mandadapu (2024), and Rebbouj and Lotfi (n.d.), I have chosen the following machine learning algorithms:

- Random Forest
- Support Vector Machines
- ElasticNet.

The reasoning behind choosing these algorithms is their widespread use in the field. Random Forests have consistently been reported as one of the best-performing models for predicting football outcomes (Baboota & Kaur, 2019; Hucaljuk & Rakipović, 2011; Rodrigues & Pinto, 2022). Using multiclass classification helps to take observations from a non-regression point of view. Support Vector Machine models have been introduced as a powerful nonlinear forecasting method (Makropoulou & Markellos, 2011; Mandadapu, 2024). Additionally, ElasticNet is preferred by Rebbouj and Lotfi (n.d.) for making predictions in football, representing a classical regression approach.

To predict the number of goals scored by each team separately, I need to set the target variables to:

1. Home Goals
2. Away Goals.

After generating the output values, I will use the Poisson distribution to calculate the probabilities of the outcomes, following the PARX method. This approach involves using the model outcomes as the lambda values, which will allow for calculating each team's goal distribution. Subsequently, the probability of exceeding or not exceeding 2.5 goals can be determined (Angelini & De Angelis, 2017).

To find the best-performing strategy, I will need to examine all possible pairs of models (3^2=9 combinations) and choose the one that yields the highest ROI. For instance, it might be that a strategy using a Random Forest for predicting Home Goals and a Support Vector Machine for predicting Away Goals yields the highest ROI. I need This leads to my first sub-question:

SQ1 - *Which two models yield the highest ROI with regards to predicting the Home and Away Goals?*

The reason for looking at two models is that each strategy combines two models; thus, the strategy yielding the highest ROI also includes two models. After comparing the strategies and selecting the best-performing pair of models based on ROI, I will analyze the most important variables and examine the construction of the chosen models. This analysis provides valuable insights into which factors significantly impact the performance of the best-performing betting strategy. By understanding the importance of these variables, we can gain a deeper understanding of how the betting strategy is constructed and which factors contribute most to its success. Additionally, examining the most important variables allows us to identify critical performance drivers and potential improvement areas. Overall, this analysis enhances our understanding of the underlying mechanisms driving the effectiveness of the betting strategy.

SQ2 - What are the most important variables for predicting the number of goals in the best-performing pair of models?

Feature selection is a crucial aspect of every method. Based on Constantinou (2019), Li and Hong (2022), Mead et al. (2023), and Saraiva et al. (2016), I expect important variables to include Home and Away advantage, Goal Count, League Points, Goal Difference, and Team Form. The pre-processing steps will discuss feature selection and data augmentation, allowing me to revisit the selected variables/features and evaluate their importance.

I will analyze the coefficients' magnitudes to visualize and discuss how each variable contributes to the predictions. Additionally, I will evaluate the average predictions for Home and Away Goals to assess the influence of the home advantage.

Some games have lower stakes as the seasons close because the final league positions might already be determined. Consequently, teams might be less motivated to perform at the same level compared to games with higher stakes (Olthof et al., 2019; Rodrigues & Pinto, 2022). Therefore, I will conduct an error analysis to examine how the ROI changes when end-of-season games are included in my dataset.

SQ3 - How do the end-season games impact the ROI?

In the end-season analysis, I will work with a smaller dataset due to the limited availability of data, focusing exclusively on the games played during the last five fixtures of each season. I will employ the two best-performing models to create a strategy and evaluate whether the ROI changes significantly. This analysis is a crucial part of error analysis, as it helps to identify and understand any discrepancies in model performance during lower-stakes games, which are prevalent at the end of the season. By isolating these matches, we can determine if their inclusion or exclusion materially affects the profitability of the models. This approach provides insights into the dynamics of end-season performance in football betting, highlighting potential weaknesses in the models and offering opportunities for refinement.

In this paper, I will adhere to the exact order of the research sub-questions because they both chronologically and logically follow each other. This means that I will first report the results of the betting strategies. Subsequently, I will evaluate the performance of the best two models. Finally, I will introduce the strategy for the end-of-season games and report its results, as it depends on the answers provided to the previous sub-questions. This structured approach provides a more straightforward and logical progression of the research findings.

4 METHODOLOGY

4.1 *Dataset and Software*

The dataset used in this study is from Kaggle and encompasses nearly 20 years of English Premier League data. It is publicly accessible (https://www.kaggle.com/datasets/louischen7/football-results-and-betting-odds-data-of-epl?select=2003-2004.csv) and was last updated three years ago. The dataset is provided in CSV format and is 1 MB in size. It is divided into Home Team and Away Team data, including separate Home Goals and Away Goals records. This dataset includes several relevant variables discussed in detail in the second sub-question (SQ2).

Here are the useful variables in the dataset:

- Full-time goals scored (Home/Away) – FTHG and FTAG in the dataset.
- Goals scored (H/A) – HTGS and ATGS.
- Goals conceded (H/A) – HTGC and ATGC.
- Team points (H/A) – HTP and ATP.
- Form points (H/A) – HTFormPts and ATFormPts.
- Match week – MW.
- Goal Difference (H/A) – HTGD and ATGD.
- Differences in points – DiffPts.
- Differences in form points – DiffFormPts.
- 3 game win streak (H/A) – HTWinStreak3 and ATWinStreak3.
- 5 game win streak (H/A) – HTWinStreak5 and ATWinStreak5.
- 3 game losing streak (H/A) – HTLossStreak3 and ATLossStreak3.
- 5 game losing streak (H/A) – HTLossStreak5 and ATLossStreak5.
- Bet365 odds (Home/Draw/Away) – B365H, B365D and B365A.
- InterWetten odds (Home/Draw/Away) – IWH, IWD, and IWA.
- Ladbrokes Coral odds (Home/Draw/Away) – LBH, LBD, and LBA.
- William Hill odds (Home/Draw/Away) – WHH, WHD, and WHA.
- Over2.5 odds – O2.5.

- Under2.5 odds – U2.5.

We can categorize the variables into five groups:

1. Traditional data: This includes variables like Goals Scored, Goals Conceded, Goal Difference, Team Points, and Differences in points. These variables are easy to interpret, representing the historical part of the dataset. They are meaningful for understanding their impact on the target variables.

2. Data about form: Form points and Differences in form points fall into this category. These variables capture recent trends or forms, emphasizing more recent data.

3. Streaks: Streaks are dummy variables in this dataset, one-hot encoded, with a value of 1 if the streak exists and 0 otherwise. These variables also capture recent trends in the data.

4. Odds: Odds are a crucial part of any dataset aiming to predict outcomes, as they contain information about the strength of the teams. A key question is how much information the odds can capture regarding the number of goals. The dataset includes home win/draw/away win odds from three bookmakers: Bet365, Interwetten, and William Hill. It is important to note that not all variables related to booking odds will be part of the model. Specifically, the over 2.5 and under 2.5 goal odds added from seasonal data are not included in the merged dataset.

5. Variables out of use: Variables not used in the model include all variables related to previous match outcomes (e.g., HM1, HM2, HM3) and league positions (HomeTeamLP, AwayTeamLP). These variables are excluded to prevent data leakage.

Since the dataset is divided into home and away data, the well-known home advantage phenomenon in football is inherently accounted for within the data's structure. Thus, my analysis does not need to address this factor separately.

The English Premier League dataset is widely used for testing betting strategies. It includes crucial variables such as the Bookmakers' Odds for Home Win, Draw, Away Win, over 2.5 goals, and under 2.5 goals, which are essential for developing my betting strategy. This dataset provides the necessary information to begin my analysis.

For my analysis, I have chosen the programming language R due to its simplicity and efficiency in handling statistical computations and data analysis. The following R packages will be utilized: caret, randomForest, Metrics, e1071, readxl, pROC, ggplot2, writexl, and MASS. These packages will enable me to effectively perform model comparison, error analysis, variable importance analysis, and cross-validation.

4.2 *Cross Validation*

I will split the data into training and test sets. The experiments will be conducted on the training set, while the test set will be used to evaluate the model. Some researchers select one or two seasons as a test set (Baboota & Kaur, 2019; Rodrigues & Pinto, 2022). Following this approach, I will use the seasons from 2002/2003 to 2016/2017 as training sets and the 2017/18 and 2018/19 seasons as test sets. This results in 15 seasons for training and 1.5 seasons for testing.

I will perform a 7-fold rolling-window cross-validation for each algorithm. The SVM and ElasticNet models will also require validation sets. Thus, 10% of each fold will be considered a validation fold, allowing for hyperparameter tuning of the SVMs and ElasticNets. Hyperparameters will be tuned directly on the training set for the RF model.

For time series datasets, rolling-origin evaluations can improve the reliability and efficiency of out-of-sample tests. Rolling-origin evaluations progressively increase the size of the training (and validation) sets with each fold as more data is collected. This technique is a reliable cross-validation method for model evaluation in time series data contexts (Tashman, 2000).

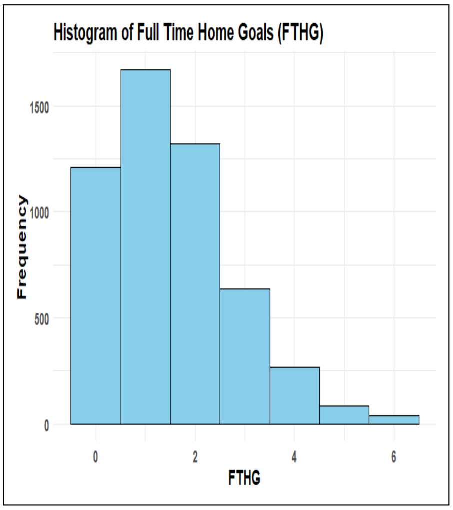
The data is progressively aggregated into folds, starting with the seasons 2002/03 and 2003/04 and continuously incorporating additional ones, such as 2004/05 and 2005/06, until the entire training set. This process results in seven folds for each model. Notably, ten percent of each fold is a validation set for SVM and ElasticNet models.

While data dynamics are not presumed to change over time, rolling-origin cross-validation can be sensitive to evolving trends. Another consideration is the presence of overlapping data attributes. Variance is likely to be higher in initial folds and decrease in later ones, potentially leading to skewed conclusions and increased bias in subsequent folds.

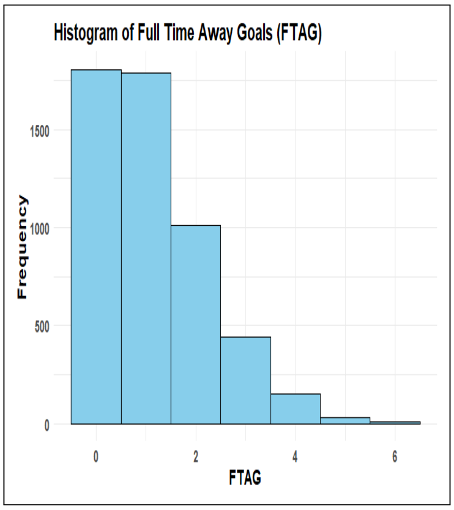
4.3 *Exploratory Data Analysis*

After downloading the dataset, I examined the customized version containing odds and removed unnecessary columns, including Date, Home Team, Away Team, and previously played game outcomes. After merging the columns, I added over and under 2.5 odds to the dataset and addressed four missing values through multiple imputations.

Upon examining the distributions of two target variables, it is evident that there exists a significant disparity between the number of goals scored at home and away. On average, teams score more goals at Home (1.52) than Away (1.14), as illustrated in Figure 1.



(a) Distributions of the Target Variables



(b) Distribution of Home Goals

Figure 1: Distribution of Away Goals

The distribution findings are of significant importance as they reveal that the most frequent number of goals scored by the Home Team is 1, whereas for the Away Team, the most frequent number is 0. This crucial information, accounting for 32% and 34.5% of the cases respectively, paints a clear picture of the scoring patterns. The second most common goal counts are 2 for the Home Team and 1 for the Away Team, further enriching our understanding of the game dynamics.

The heatmap of the correlation matrix, excluding only the B365 odds for more precise visualization, reveals the strong correlations among them, as depicted in Figure 2. Among the independent variables, B365 odds demonstrate the strongest correlations in the heatmap. A negative correlation exists between Home Goals and B365H odds (-0.32) and Away Goals and B365A odds (-0.29). Conversely, positive correlations are observed between Home Goals and B365A odds (0.36) and Away Goals and B365H odds (0.34). These relationships underscore the significant influence of betting odds on goal outcomes, an essential aspect of our analysis.

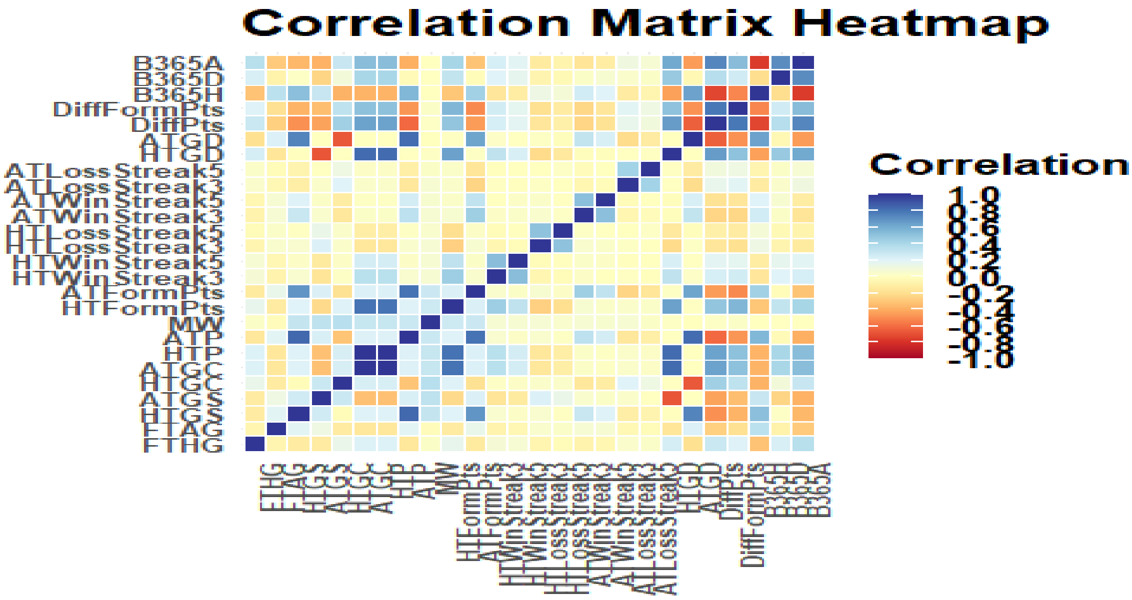


Figure 2: Heatmap of correlation matrix

It's important to note that certain variables in the dataset are transformed to aid interpretation. These transformations, such as dividing them by the match week variable, are not just for statistical purposes but also practical reasons. This transformation is a form of normalization, enhancing consistency and making the variables more comparable across seasons. Variables such as Goal differences, Team points, Differences in points, and Differences in form points are initially kept in this format. I also found it reasonable to transform goals scored and conceded for the same reasons, as it allows for a more accurate and meaningful data analysis with direct implications for football prediction and betting strategies.

5 Preprocessing

After these adjustments, I separated the end-of-season games from the rest and saved them in another file, all in Excel format. This process involved identifying the last five fixtures of each season (match week 34-38) and distinguishing them from the rest of the dataset. The decision to draw the line at this point is grounded in the observation that, typically, five fixtures before the end of the season, many uncertainties still exist regarding relegation battles, title races, or positions in the Champions League and Europa League. Theoretically, teams should be 15 points clear at this stage, but this is hardly the case for most teams.

After merging the datasets and conducting multiple imputations, I proceeded with an outlier analysis. Initially, I employed the Mahalanobis distance, which yielded satisfactory results. Following this, I applied Tukey's boxplot method. Any value falling outside the outer bounds was replaced with the closest non-outlier value. Subsequently, I standardized all predictor values except for the dummy variables.

The target variables, Home Goals and Away Goals have integer values ranging from 0 to 9. However, after applying Tukey's boxplot method and the 7-fold cross-validation, I had to set the maximum value of the target variables to 6. This decision was based on the analyses conducted to mitigate the potential influence of outliers. Outliers have the potential to introduce noise, increase the variance of the residuals, or adversely affect the bias. This approach also allows uniform labeling of Home Goals and Away Goals across the dataset, facilitating a coherent analysis.

5.1 *Machine Learning in General*

Machine learning techniques have increasingly been applied to football prediction tasks. Over a decade ago, Bayesian Networks and regression models yielded impressive results. However, recent research highlights the growing efficacy of random forests, support vector machines (SVMs), k-nearest neighbors (KNNs), and gradient-boosting methods. It has been demonstrated that successful betting strategies can be developed using machine learning algorithms (Heijboer, 2022). The return on investment (ROI) for these strategies can vary significantly, ranging from -14.06% to 20.08% (Constantinou, 2019), or between 2.3% and 26.7% (Rodrigues & Pinto, 2022), depending on the machine learning tools employed. The upper bound of the Dolores Model was achieved in the English Premier League over seven consecutive seasons. The methods I have chosen—Random Forest, Support Vector Machine, and ElasticNet—are widely used for making predictions in the football domain.

As you will see, I have chosen three regression techniques to build my models. I have chosen regression techniques over classification to avoid limiting predictions to specific categories. This approach allows for more precise and nuanced predictions, preserving the ordinal nature of the number of goals scored, which would be lost in a classification task.

5.2 *Random Forest*

Random Forests are an ensemble technique known for their outstanding performance, particularly with large datasets and those containing numerous variables (Dahinden & Ethz, 2011). One limitation of Random Forests is their complexity and computational intensity, but they remain widely used in football prediction tasks. When building a Random Forest model, I will focus on two hyperparameters: the number of trees and randomly selected variables as candidates per split (mtry). There are two types of Random Forests (RF): the classical RF for classification tasks and RF Regression for regression tasks. Common metrics for evaluating RF Regression models include R^2, Mean Absolute Error (MAE), and Mean Squared Error (MSE). Since Random Forests handle regression tasks, the target variable must be numeric. To ensure consistency across all folds and the test set, the upper bound of the target variables must be set to 6.

5.3 *Support Vector Machine*

Support Vector Machines (SVMs) can be categorized into two types: the classical SVM for classification tasks and the Support Vector Regression (SVR) for regression tasks. The classical SVM aims to find the optimal hyperplane to separate classes. In contrast, SVR aims to find a function or equation that estimates the relationship between predictor variables and the target variable.

One of the advantages of SVMs is their applicability to both linear and non-linear tasks. They are highly interpretable, suitable for regression and classification problems, and can be applied to time series data. SVMs involve using a radial kernel and tuning the gamma parameter. The key hyperparameters for the SVM model include the kernel type, cost value, gamma value, and epsilon value. Tuning these parameters can incur computational costs due to their sensitivity to settings (Igiri, 2015).

5.4 *ElasticNet*

ElasticNet is a form of regularized optimization, providing an interpretable and classical approach for regression analysis. In the ElasticNet model, two parameters will be tuned: lambda and alpha. The lambda parameter controls the balance between Lasso (L1) and Ridge (L2) regularization, while the alpha parameter determines the overall strength of regularization. Higher values of alpha lead to greater shrinkage of the coefficients towards zero, enhancing model simplicity and reducing overfitting (Hans, 2011).

ElasticNet might have a high bias with the non-linear dummy predictors as a linear model, potentially leading to underfitting. To mitigate this, I will scale the variables by standardizing them. Another limitation of ElasticNet is that shrinking the coefficients close to zero can make it difficult to interpret the model's coefficients.

5.5 *Evaluation Metrics and ROI*

Since I work with three regression-based models, I evaluate all the models based on R^2 and Mean Absolute Error (MAE). R^2 quantifies the proportion of variance in the model explained by the dependent variables, providing insight into how well SVM and ElasticNet fit the data. MAE is preferred over Mean Squared Error (MSE) due to its interpretability; MAE is expressed in the original units of the dataset, whereas MSE returns squared units. Moreover, MAE treats all errors equally, making it less sensitive to outliers.

Return on Investment (ROI) is a crucial financial metric that measures the efficiency of an investment relative to its cost (Friedlob & Plewa Jr, 1996). In football betting, ROI is the ratio of a bettor's net gain to the total bet amount placed (Ethier & Ethier, 2010). Unlike simple profit, ROI accounts for the factor of time, providing a comprehensive evaluation of strategy performance over the long term. By emphasizing ROI, I ensure a robust assessment of strategy effectiveness beyond traditional model evaluation metrics.

5.6 *Flowchart*

Below, I provide the flowchart of my work.

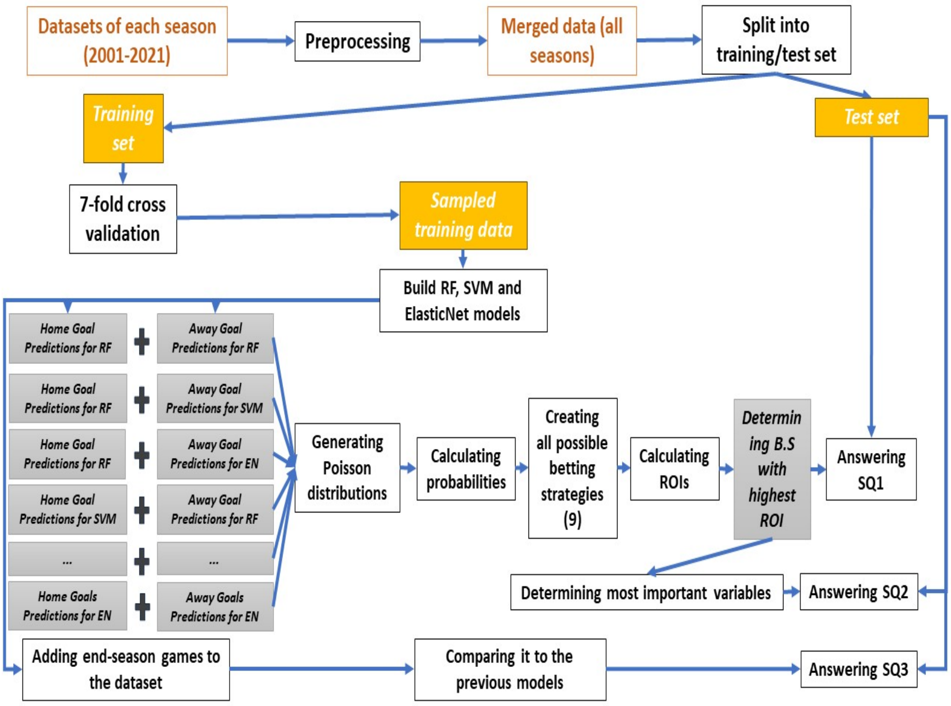


Figure 3: Thesis flowchart

5.7 *Hyperparameter Tuning and Model Evaluation*

Algorithm	Target variable	Model parameters	Performance
RF	HG	mtry_value <- 15 ntree_value <- 9	Test R-squared: 0.130139 Test Mean Absolute Error (MAE): 0.9863248
	AG	mtry_value <- 33 ntree_value <- 9	Test R-squared: 0.1156161 Test Mean Absolute Error (MAE): 0.8708577
SVM	HG	cost_value <- 1 gamma_value <- 0.001 epsilon_value <- 0.001	Test R-squared: 0.1726544 Test Mean Absolute Error (MAE): 0.935678
	AG	cost_value <- 1 gamma_value <- 0.001 epsilon_value <- 0.001	Test R-squared: 0.172986 Test Mean Absolute Error (MAE): 0.8203191
EN	HG	alpha = 0.4 lambda = 0.4	Test R-squared: 0.170844 Test Mean Absolute Error (MAE): 1.108553
	AG	alpha = 0.2 lambda = 0.8	Test R-squared: 0.1952623 Test Mean Absolute Error (MAE): 0.8405845

Table 1: Performance Summary

As observed, both Random Forest (RF) models perform best with nine trees. This setup prevents underfitting, which can occur when the number of trees is too low. However, RF regression models are among the worst-performing models compared to the other techniques.

For the Home Goal (HG) model, the mtry value, representing the number of randomly selected variables for each tree, was set to 15. This reduces model complexity, which is necessary given the high number of trees. In contrast, the mtry value for the Away Goals (AG) model was set to 33, the maximum capacity given the 33 predictor variables.

During hyperparameter tuning of the SVM models, I set both to the same parameters. I selected the linear kernel as it outperformed the radial kernel. The cost value, a sensitive parameter, was adjusted for optimal performance. Both gamma and epsilon values were set to 0.001, where a low epsilon value indicates a stricter fit to the data, and the gamma value results in a smoother decision boundary.

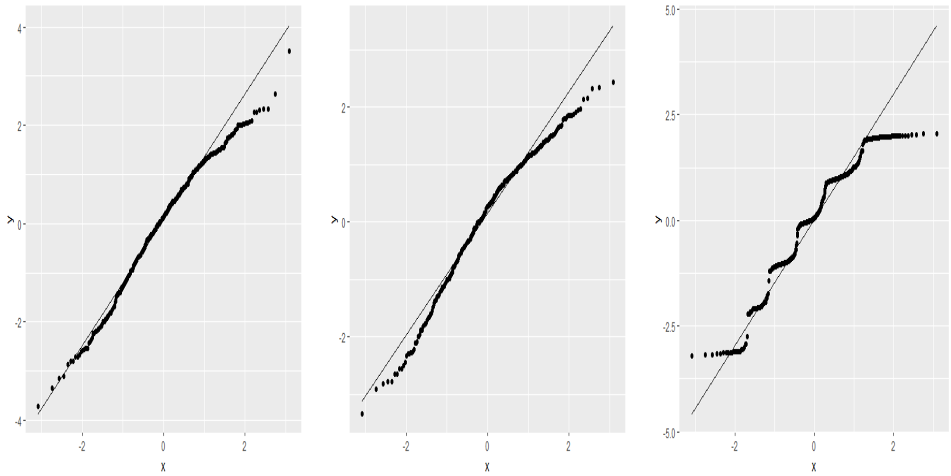
For ElasticNet models, the alpha and lambda values differ between the two models. The HG model has an alpha of 0.4, closer to Ridge regression with moderate regularization (lambda = 0.4). With a stronger regularization (alpha = 0.2 and lambda = 0.8), the AG model shrinks the coefficients significantly, although not to zero.

The R^2 value for the Home Goal (HG) model is 0.171, while the Away Goal (AG) model has an R^2 value of 0.195. These R^2 values are not considered exceptional in this field. The Mean Absolute Error (MAE) for the HG model is significantly higher than for the AG model, likely due to differences in standard deviation between home and away goals. This trend is consistent across the models.

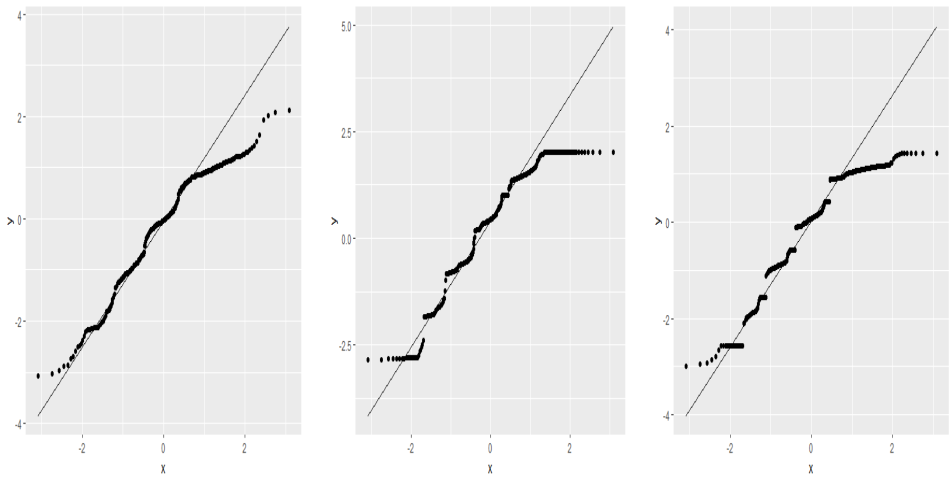
The R^2 values of the models range between 11% and 20%, which is typical in this field. The MAEs of the AG models are consistently lower than those of the HG models, likely due to the higher standard deviation of home goals (1.29 for home goals vs. 1.11 for away goals).

The RF models are the worst performers based on R^2 and MAE. The ElasticNet model for AG has the highest R^2, while the SVM model for AG has the lowest MAE. Therefore, I expect the best-performing strategies to include either the ElasticNet or SVM models for AG, while RF models are unlikely to be part of any top-performing strategies.

5.8 *Error Analysis*



(a) Random Forest HG (b) Random Forest AG (c) SVM HG



(d) SVM AG (e) ElasticNet HG (f) ElasticNet AG

Figure 4: QQ plots

I presented QQ plots for each model. In the QQ plots of the Random Forest models, the observations predominantly fall below the reference line, with some points occasionally touching or slightly exceeding it. In contrast, SVM models display discernible patterns characterized by a less concentrated distribution than ElasticNet models, which exhibit a more tightly clustered arrangement. Both SVM and ElasticNet models demonstrate extended tails, indicating deviations from normality towards the upper and lower quantiles.

None of the QQ plots indicate conformity to a normal distribution, which aligns with the target variables' non-normal distribution. Therefore, the observed patterns in the QQ plots are not unexpected or problematic for my analysis.

5.9 *Creating Probabilities*

After making predictions on the test set, I created probabilities, as shown in the PARX model. This process can also be accomplished in R using the packages mentioned earlier.

The first step in creating the probability distributions is determining the expected number of goals for the home and away teams (generating predictions). Let's take the first prediction of the SVM HG model as an example: 1.259. According to the SVM algorithm, this value represents the expected number of goals for the home team.

The next step is to create a Poisson distribution using this lambda value. Here, lambda equals 1.259 (the expected number of goals), and the number of events (x) is 0, 1, 2, 3, 4, 5, and 6. With these parameters, we can derive the probability distribution for the home team's goal count, as shown in Table 2. It's important to note that the likelihood of scoring six goals is calculated by subtracting the cumulative probability of all other possible outcomes from 1.

x	**Probability with** $\lambda = 1.259$
0	0.283937822
1	0.357477718
2	0.225032224
3	0.094438523
4	0.029724525
5	0.007484635
6	0.001904552

Table 2: Probability Distribution with $\lambda = 1.259$

This process can also be applied to the AG models, resulting in two distributions: one for the home team's goals and one for the away team's goals for each algorithm.

Next, I needed to determine the probability of scoring over and under 2.5 goals in a game. Calculating the likelihood of each match result is straightforward after multiplying the appropriate events in the two distributions. By focusing on specific match results such as 0-0, 0-1, 1-0, 1-1, 0-2, and 2-0, we can determine the probability of scoring under 2.5 goals by summing the probabilities associated with these outcomes. Consequently, calculating the probability of scoring over 2.5 goals is complimentary:

$$P(\text{over } 2.5) = 1 - P(\text{under } 2.5)$$

After this step, determining the higher probability and selecting the preferred outcome for betting becomes straightforward.

5.10 *Betting Strategies*

Following the creation of betting strategies, the next step is to finalize and calculate the profits and the return on investments (ROI) of these strategies (Ethier & Ethier, 2010). Nine possible betting strategies can be created by combining the following models:

1. Random Forest model for Home Goals (RF HG model) with Random Forest model for Away Goals (RF AG model)
2. RF HG model with Support Vector Machine model for Away Goals (SVM AG model)
3. RF HG model with ElasticNet model for Away Goals (ElasticNet AG model)
4. Support Vector Machine model for Home Goals (SVM HG model) with RF AG model
5. SVM HG model with SVM AG model
6. SVM HG model with ElasticNet AG model
7. ElasticNet model for Home Goals (ElasticNet HG model) with RF AG model
8. ElasticNet HG model with SVM AG model
9. ElasticNet HG model with ElasticNet AG model

To compare the strategies, we need to evaluate the profits and the ROIs, which depend on the accuracy of the bets. Let's assume a stake of 1 € per game. The initial bankroll should be 494 €, corresponding to the size of the final test set. If a bet is incorrect, the loss is 1 €. Conversely, if a bet is correct, the profit is calculated as follows: $1 \times \text{bet}_{\text{odds}} - 1$. For example, if a game with odds of 1.82 is predicted correctly, the profit would be $1 \times 1.82 - 1 = 0.82$. This method allows for a straightforward comparison of the effectiveness and profitability of each betting strategy (Rodrigues & Pinto, 2022).

5.11 *Betbrain Max and Betbrain Average*

The dataset contains two types of odds from BetBrain (BB), an online platform that allows bettors to compare offers from different bookmakers to make the most profitable bets. The BB Average odds represent the average of the odds across all bookmakers, while the BB Max odds are the highest offered by any bookmaker. Consequently, the BetBrain Max odds are better than the averaged ones.

I will evaluate the strategies' results using BB Max and BB Average odds. This approach allows me to compare the difference between betting on the best and average offers available in the market.

6 RESULTS

This paper has two results sections: one for the results of betting strategies and one for the results of end-season games. To maintain chronological order, I will report the results in two sections and make conclusions in one final section.

6.1 *Results of Betting Strategies*

Strategy	**Profit - BB Average (€)**	**Profit - BB Max (€)**	**ROI (Average)**	**ROI (Max)**
Strategy_RF_HG_RF_AG	-18.47	-0.58	-3.74%	-0.12%
Strategy_RF_HG_SVM_AG	-30.92	-13.42	-6.26%	-2.72%
Strategy_RF_HG_EN_AG	-38.93	-21.05	-7.88%	-4.26%
Strategy_SVM_HG_RF_AG	-7.1	11.96	-1.44%	2.42%
Strategy_SVM_HG_SVM_AG	-47.79	-30.45	-9.67%	-6.16%
Strategy_SVM_HG_EN_AG	-67.73	-51.22	-13.71%	-10.37%
Strategy_EN_HG_RF_AG	4.76	24.06	0.96%	4.87%
Strategy_EN_HG_SVM_AG	-28.54	-11.09	-5.78%	-2.24%
Strategy_EN_HG_EN_AG	-63.78	-47.16	-12.91%	-9.55%

Table 3: Profits and the Bettor's ROI

The results demonstrate consistent and coherent outcomes across various strategy categories. Surprisingly, the combination of ElasticNet for Home Goals and Random Forest for Away Goals consistently achieves the highest ROI across all tested strategies. This approach yielded a profit of €4.76 with an ROI of 0.96% when utilizing the best bets from BetBrain. Combining the Support Vector Machine for HG and RF for AG models proved profitable using BetBrain's maximum odds.

Conversely, the least effective strategy involved SVM for HG and ElasticNet for AG models, resulting in a substantial loss of €67.73 with an ROI of -13.71% using the averaging method. Interestingly, three out of four worst-performing strategies included the ElasticNet AG model, indicating its limited effectiveness in this context.

Notably, only one strategy showed profitability using the averaging method alone, suggesting that superior profits are achievable when leveraging BetBrain's top odds. Moreover, out of the nine strategies evaluated, only two yielded a positive ROI when employing maximum odds.

Combining EN for HG and RF for AG emerges as the best-performing strategy based on ROI. Given these promising results, further feature analysis of these models will be pursued to understand their predictive power and synergy better.

6.2 *Feature Analysis of the best performing two models*

6.2.1 *EN HG model*

After examining the magnitudes of coefficients, it becomes evident that odds play a significant role in ElasticNet model construction. The top four predictors are all related to odds, reflecting their strong correlation. Away Team Goals Scored (ATGS) ranks among the top 5 most influential variables.

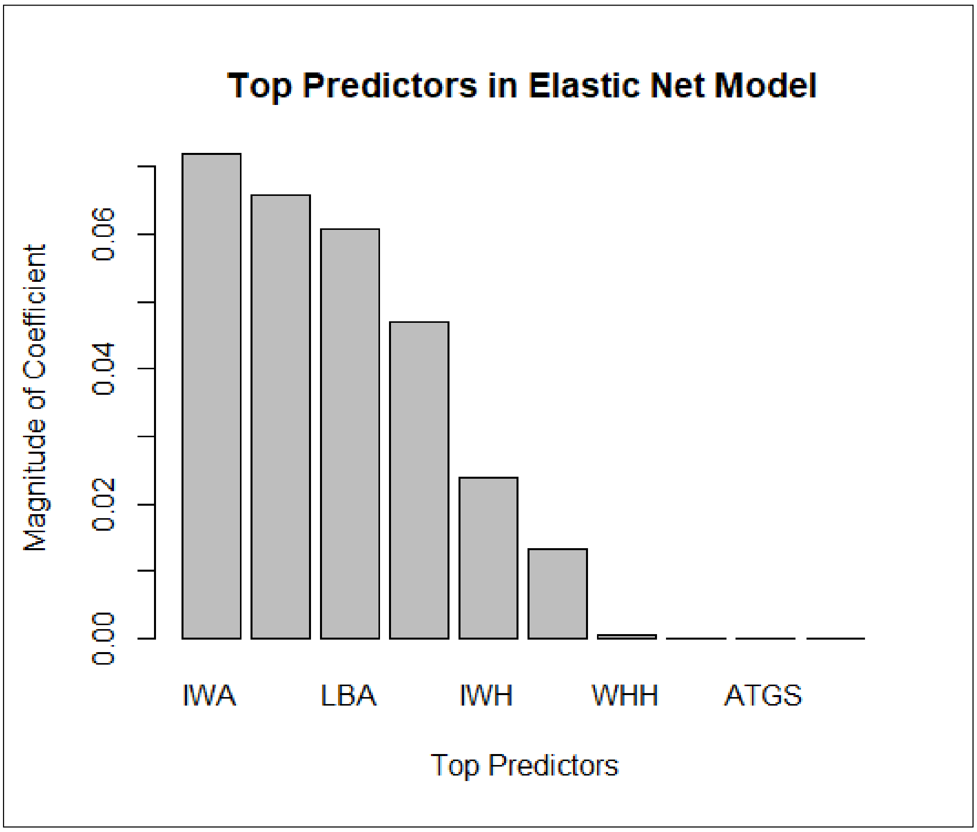


Figure 5: EN HG most important variables

6.2.2 *RF AG model*

I utilized the Mean Decrease in Impurity measure to assess each predictor's contribution to variance reduction in decision trees. Higher values indicate more significant predictive power. Unlike the ElasticNet model, predictors are more evenly distributed here, with only one odd feature among the top 3. This suggests that odds are less dominant in Random Forest models than ElasticNets. Instead, traditional features such as Home Team Goals Scored (HTGS) and Away Team Points (ATP) are slightly more influential than those containing the odds.

Although neither model explicitly incorporates home and away advantages in variable analysis, both models account for these factors by separately predicting home and away goals. This effect is best illustrated through prediction averages: the EN HG model averages 1.52 for home goals, whereas the RF AG model averages 1.29. This indicates that, on average, predictions for away goals are lower than those for home goals.

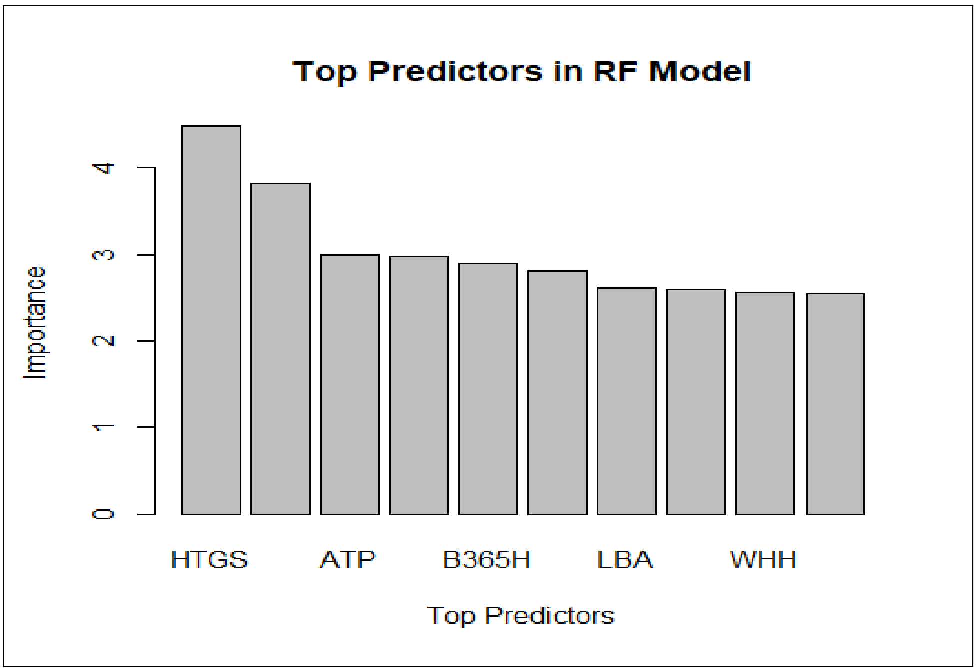


Figure 6: RF AG most important variables

6.3 *End season analysis*

Conducting the end-season analysis differs slightly from the previous section. Here, I focus solely on the changes made compared to earlier models.

Firstly, it's important to note that the dataset is more petite, covering match weeks 34 to 38. Due to missing end-season games, I have data for only 15 seasons, totaling 719 observations. Consequently, the training set includes 13 end-season games, while the test set comprises data from the 2016/17 and 2017/18 seasons, maintaining previous ratios.

I implemented a 4-fold rolling-origin cross-validation with validation sets. The Random Forest (RF) model used a lower mtry value to optimize performance.

Algorithm	Target variable	Model parameters	Performance
EN	HG	lambda_value <- 0.4 alpha_value <- 0.4	Test R-squared: 0.2069053 Test Mean Absolute Error (MAE): 1.106127
RF	AG	mtry_value <- 20 ntree_value <- 9	Test R-squared: 0.1005539 Test Mean Absolute Error (MAE): 0.9434815

Table 4: End Season Model - Hyperparameters

6.4 *Results of the end season analysis*

Evaluating the performance of the best-performing model followed the same approach as previously. The test set comprised 100 observations, revealing negative results. Using the BetBrain Max odds, there was a -60.37% ROI, and with averaged odds, the ROI was 62.21%.

Strategy	Profit - BB Average (€)	Profit - BB Max (€)	ROI (Average)	ROI (Max)
EN_HG_RF_AG	-61.21	-59.37	-62.21%	-60.37%

Table 5: Performance of the End Season Strategy

7 DISCUSSION

7.1 *To what extent can a successful betting strategy be created using Machine Learning Models?*

The negative ROI results are unsurprising given the models' performance metrics and the average winning odds, which are 1.89 (average) and 1.97 (maximum). Achieving high profits in the betting market typically requires models to perform exceptionally well because the average winning odds are not very high. This observation aligns with one of the core principles of a profitable betting company.

Several factors might explain the models' performance. One significant reason could be feature selection. Predicting the number of goals is complex and requires specific variables to create a well-performing model. Important variables might include the team's average expenditure, individual player values, player ages, or ELO scores (Mead et al., 2023).

Another potential issue could be the lack of experimentation with training and test data sizes. Research has shown that data splits can significantly impact a model's performance in football predictions (Mandadapu, 2024).

7.2 *Which two models yield the highest ROI with regards to predicting the Home and Away Goals?*

Generally speaking, most models yielded a negative ROI. However, two strategies produced a positive ROI using Betbrain's best offers. Notably, both profitable strategies included the Random Forest (RF) model for predicting Away Goals (AG). Combining an ElasticNet model for predicting Home Goals (HG) and a Random Forest model for predicting Away Goals yielded the highest ROI.

7.3 *SQ2 - What are the most important variables for predicting the number of goals in the best-performing pair of models?*

The odds had the most significant impact on the ElasticNet HG model. Specifically, IWA, LBA, IWH, and WHH were the most important, as their coefficients had the highest magnitudes. These top four predictors accounted for most of the explained variance, but the Away Team Goals Scored coefficient was also among the most influential.

In the Random Forest AG model, all variables were included in each decision tree, with the model comprising nine decision trees. The Mean Decrease in Impurity measure indicated that the features were more evenly distributed, meaning no single feature dominated in importance. However, Home Team Goals Scored, Away Team Points, and the odds were among the most essential variables.

Notably, the home advantage is evident in the findings. The ElasticNet Home Goals (EN HG) model predicts an average of 1.52 home goals, while the Random Forest Away Goals (RF AG) model predicts an average of 1.29 away goals. This discrepancy indicates the presence of a home advantage.

7.4 *SQ3 - How do the end-season games impact the ROI?*

Recreating the best strategy based on end-season data led to a significant decrease in ROI. This decline can be attributed to several factors. One reason might be that end-season games have lower stakes, causing teams not to perform at the same level as during mid-season. Another reason could be the lack of data, as the test set contained only 100 observations, which might not be representative.

7.5 *Limitations*

The dataset did not contain data from 2019, so I could not check the most recent trends. Additionally, the test size of the end-season games is small, making the results less robust. Consequently, this aspect of my work may generate hypotheses rather than definitive conclusions.

I used three algorithms to make predictions on two target variables, but some more algorithms and strategies could be tested. For instance, a multi-class classification might be more profitable than creating a strategy using a classical Support Vector Machine instead of Support Vector Regression.

In this study, I attempted to create predictions indirectly by first predicting the number of goals for the home and away teams. This approach presents a limitation, as it complicates the task by combining multiple methods, making it more challenging to distinguish the impact of each technique. This limitation arises directly from my chosen approach to handling the task, potentially affecting the overall effectiveness of the predictions.

Furthermore, I did not include any interactions or information about the odds from BetBrain. Adding more variables and interactions to the models could improve performance and increase the ROI of the strategies.

It is also possible that using these machine learning techniques is more profitable for predicting outcomes rather than the total number of goals. This suggests that my methods may not be the most suitable for predicting the total number of goals.

In summary, the disparate impact of my research approach is evident in several aspects. The choice to predict goals indirectly through multiple methods complicates the evaluation of individual techniques. The exclusion of odds and interaction variables limits model performance. Lastly, focusing on goal totals rather than outcomes may not align with the most profitable application of machine learning techniques in football predictions. These factors collectively influence the overall effectiveness and profitability of the predictions made in this study.

7.6 *Future Work*

Collecting data and information can enhance model performance and lead to more profitable strategies, especially on specific features like distance, player ratings, ELO ratings, weather conditions, attendance, average age, and goal-scoring streaks.

Additionally, the results may be improved by directly predicting the total number of goals in each game rather than making predictions in two rounds. This direct approach would be computationally faster and simplify evaluating each method's performance.

Another potential improvement is to experiment with the probabilities of each over/under prediction to find a threshold for placing bets. This would allow distinguishing bets based on the confidence level of each prediction.

Exploring more algorithms, such as Gradient Boosting, K-Nearest Neighbors (KNN), or logistic regression, could provide additional strategies and potentially better results. Furthermore, experimenting with different training and test split ratios could enhance the model's effectiveness (Heijboer, 2022).

Lastly, targeting other events, such as the number of cards, throw-ins, or corners, presents a largely unexplored area in the betting market that could yield valuable insights and profitable strategies.

8 CONCLUSION

First, I created nine betting strategies by separately predicting the number of home and away goals using RF, SVM, and ElasticNet models and generating Poisson distributions. I calculated the probability of scoring over/under 2.5 goals in a football game. Based on the ROI of the strategies, I selected the strategy with the highest ROI and examined the most critical variables. Finally, I recreated the two best models and assessed their impact on the ROI.

The results showed that most strategies did not yield a positive return, with only two out of nine models proving profitable when using the advantageous odds from the BetBrain platform. The most profitable strategy was the combination of an ElasticNet Home Goal and a Random Forest Away Goal model, which yielded an ROI of 0.96%.

Feature importance analysis indicated that the EN HG model heavily relies on away win odds, with high magnitudes of coefficient values reflecting solid correlations among the top predictors. Conversely, the RF AG model shows a more balanced distribution of predictors, with traditional features like Home Team Goals Scored, Away Team Points, and B365H odds being the most important.

The end-season analysis revealed a significant decrease in the ROI of the best strategy, possibly because teams do not perform at the same level when their games are less important.

Regarding my previous aspirations to address the neglected area of football betting, I believe that the results and limitations of this paper highlight the expanded possibilities in the market. This can lead to more informed decisions for betting market participants. It also shows that paying attention to events other than football outcomes has the potential to create successful strategies in the future.

REFERENCES

Angelini, G., & De Angelis, L. (2017). Parx model for football match predictions. *Journal of Forecasting*, *36*(7), 795–807.

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, *35*(2), 741–755.

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*(2), 458–466.

Choi, B. S., Foo, L. K., & Chua, S.-L. (2023). Predicting football match outcomes with machine learning approaches. *MENDEL*, *29*(2), 229–236.

Constantinou, A. C. (2019). Dolores: A model that predicts football match outcomes from all over the world. *Machine learning*, *108*(1), 49–75.

Dahinden, C., & Ethz, M. (2011). An improved random forests approach with application to the performance prediction challenge datasets. *Hands-on Pattern Recognition, Challenges in Machine Learning*, *1*(2), 223–30.

Ethier, S. N., & Ethier, S. N. (2010). Optimal proportional play. *The Doctrine of Chances: Probabilistic Aspects of Gambling*, 357–390.

Friedlob, G. T., & Plewa Jr, F. J. (1996). *Understanding return on investment*. John Wiley & Sons.

Glickman, M. E., & Jones, A. C. (1999). Rating the chess rating system. *CHANCE-BERLIN THEN NEW YORK-*, *12*, 21–28.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association*, *106*(496), 1383–1393.

Heijboer, M. (2022). *Predicting football match outcomes using machine learning algorithms* [Doctoral dissertation, Tilburg University].

Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques. *2011 Proceedings of the 34th International Convention MIPRO*, 1623–1627.

Igiri, C. P. (2015). Support vector machine–based prediction system for a football match result.

Li, Y., & Hong, Y. (2022). Prediction of football match results based on edge computing and machine learning technology. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, *13*(2), 1–10.

Makropoulou, V., & Markellos, R. N. (2011). Optimal price setting in fixed-odds betting markets under information uncertainty. *Scottish Journal of Political Economy*, *58*(4), 519–536.

Mandadapu, P. (2024). The evolution of football betting-a machine learning approach to match outcome forecasting and bookmaker odds estimation. *arXiv preprint arXiv:2403.16282*.

Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *Plos one*, *18*(4), e0282295.

Olthof, S. B., Frencken, W. G., & Lemmink, K. A. (2019). When something is at stake: Differences in soccer performance in 11 vs. 11 during official matches and training games. *The Journal of Strength & Conditioning Research*, *33*(1), 167–173.

Rebbouj, M., & Lotfi, S. (n.d.). Football league ranking prediction using machine learning regression model. *Annals of Applied Sport Science*, 0–0.

Rodrigues, F., & Pinto, Â. (2022). Prediction of football match results with machine learning. *Procedia Computer Science*, *204*, 463–470.

Saraiva, E. F., Suzuki, A. K., Louzada, F., et al. (2016). Predicting football scores via poisson regression model: Applications to the national football league. *Communications for Statistical Applications and Methods*, *23*(4), 297–319.

Stübinger, J., Mangold, B., & Knoll, J. (2019). Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, *10*(1), 46.

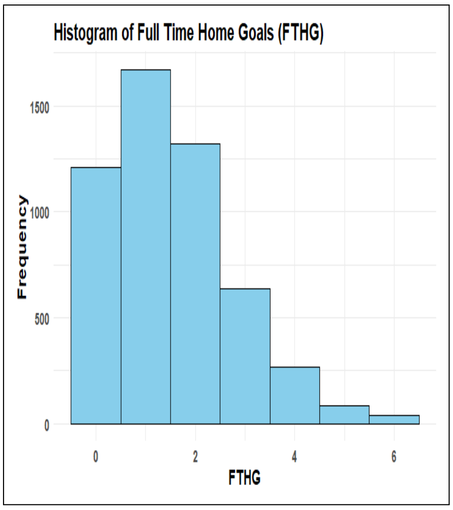
Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International journal of forecasting*, *16*(4), 437–450.

Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, *36*(3), 916–932.

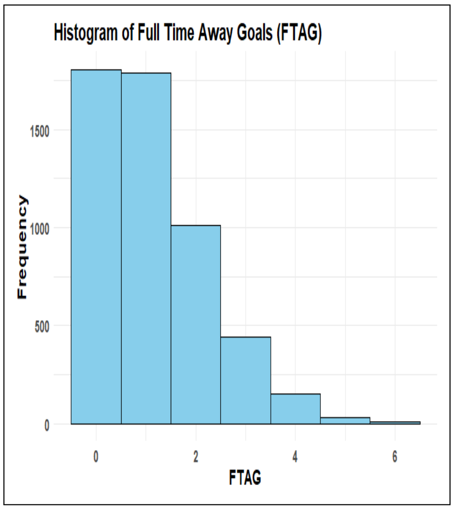
Wheatcroft, E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics*, *7*(2), 77–97.

Lixi Chen. (2022). Football Results and Betting Odds Data of English Premier League (EPL). Kaggle. https://www.kaggle.com/datasets/louischen7/football-results-and-betting-odds-data-of-epl?select=2003-2004.csv

9 APPENDIX



(a) Distributions of Away Goals



(b) Distribution of Home Goals

Figure 1: Distribution of the Target Variables

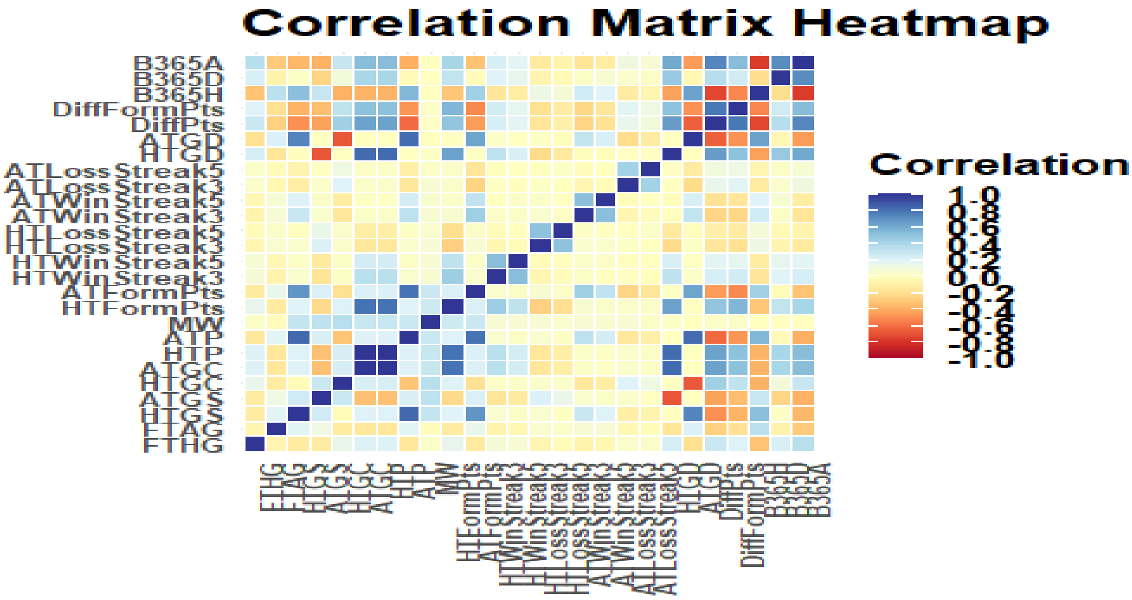


Figure 2: Heatmap of correlation matrix

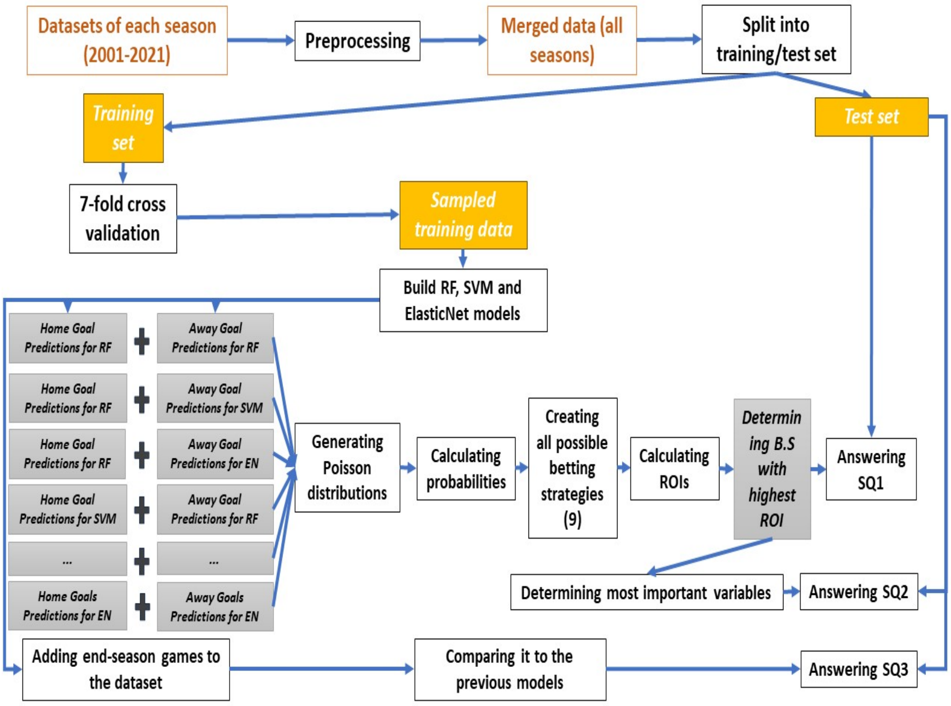
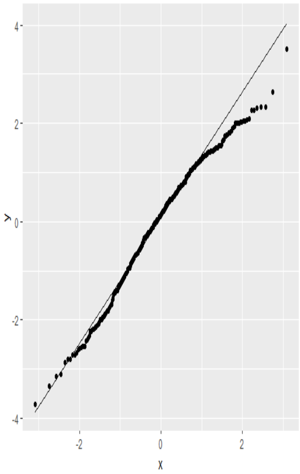


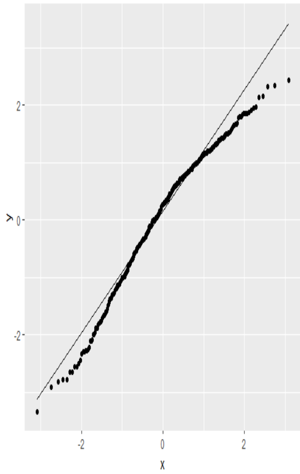
Figure 3: Thesis flowchart

Algorithm	**Target variable**	**Model parameters**	**Performance**
RF	HG	mtry_value <- 15 ntree_value <- 9	Test R-squared: 0.130139 Test Mean Absolute Error (MAE): 0.9863248
	AG	mtry_value <- 33 ntree_value <- 9	Test R-squared: 0.1156161 Test Mean Absolute Error (MAE): 0.8708577
SVM	HG	cost_value <- 1 gamma_value <- 0.001 epsilon_value <- 0.001	Test R-squared: 0.1726544 Test Mean Absolute Error (MAE): 0.935678
	AG	cost_value <- 1 gamma_value <- 0.001 epsilon_value <- 0.001	Test R-squared: 0.172986 Test Mean Absolute Error (MAE): 0.8203191
EN	HG	alpha = 0.4 lambda = 0.4	Test R-squared: 0.170844 Test Mean Absolute Error (MAE): 1.108553
	AG	alpha = 0.2 lambda = 0.8	Test R-squared: 0.1952623 Test Mean Absolute Error (MAE): 0.8405845

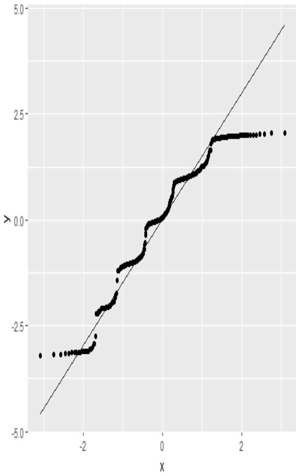
Table 1: Performance Summary



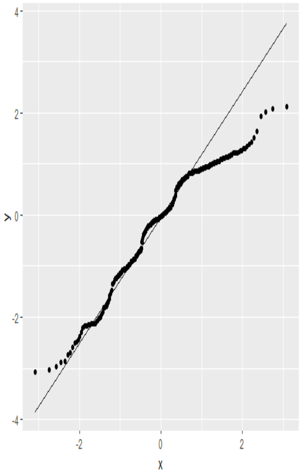
(a) Random Forest HG



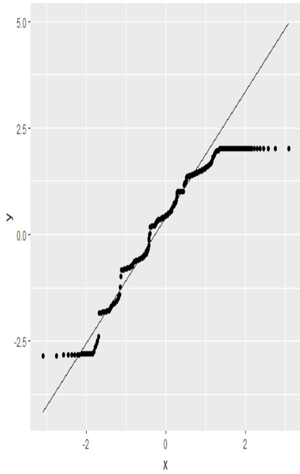
(b) Random Forest AG



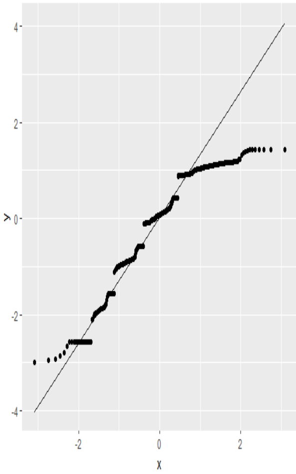
(c) SVM HG



(d) SVM AG



(e) ElasticNet HG



(f) ElasticNet AG

Figure 4: QQ plots

x	**Probability with** $\lambda = 1.259$
0	0.283937822
1	0.357477718
2	0.225032224
3	0.094438523
4	0.029724525
5	0.007484635
6	0.001904552

Table 2: Probability Distribution with $\lambda = 1.259$

Strategy	Profit - BB Average (€)	Profit - BB Max (€)	ROI (Average)	ROI (Max)
Strategy_RF_HG_RF_AG	-18.47	-0.58	-3.74%	-0.12%
Strategy_RF_HG_SVM_AG	-30.92	-13.42	-6.26%	-2.72%
Strategy_RF_HG_EN_AG	-38.93	-21.05	-7.88%	-4.26%
Strategy_SVM_HG_RF_AG	-7.1	11.96	-1.44%	2.42%
Strategy_SVM_HG_SVM_AG	-47.79	-30.45	-9.67%	-6.16%
Strategy_SVM_HG_EN_AG	-67.73	-51.22	-13.71%	-10.37%
Strategy_EN_HG_RF_AG	4.76	24.06	0.96%	4.87%
Strategy_EN_HG_SVM_AG	-28.54	-11.09	-5.78%	-2.24%
Strategy_EN_HG_EN_AG	-63.78	-47.16	-12.91%	-9.55%

Table 3: Profits and the Bettor's ROI

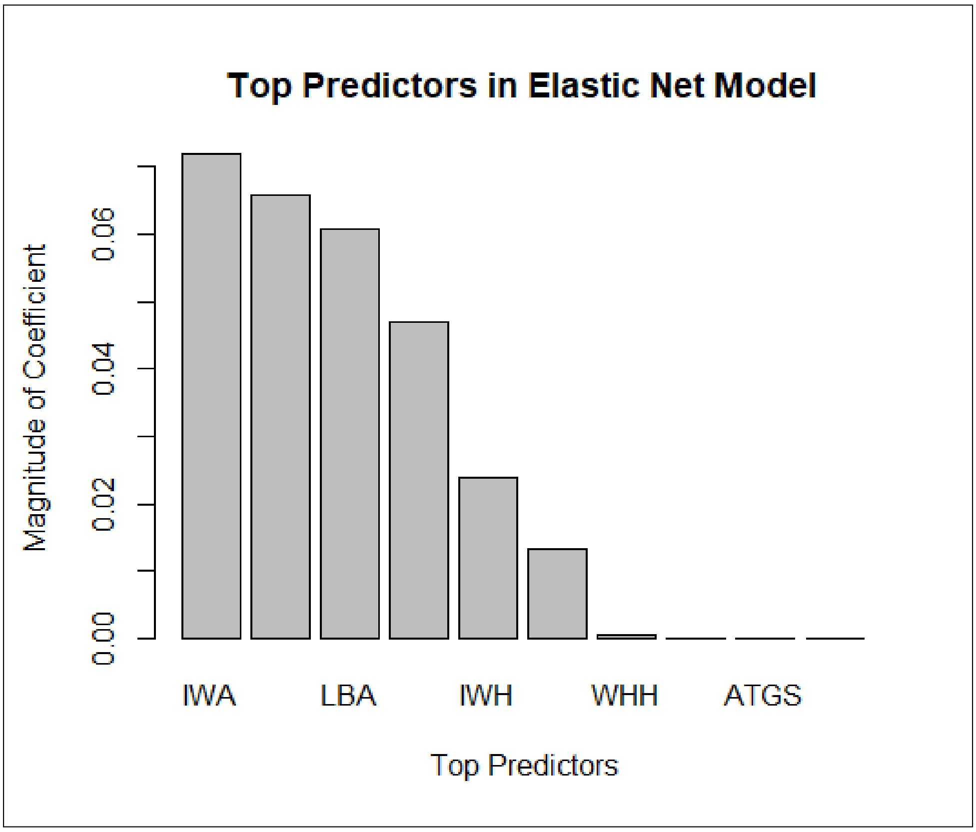


Figure 5: EN HG most important variables

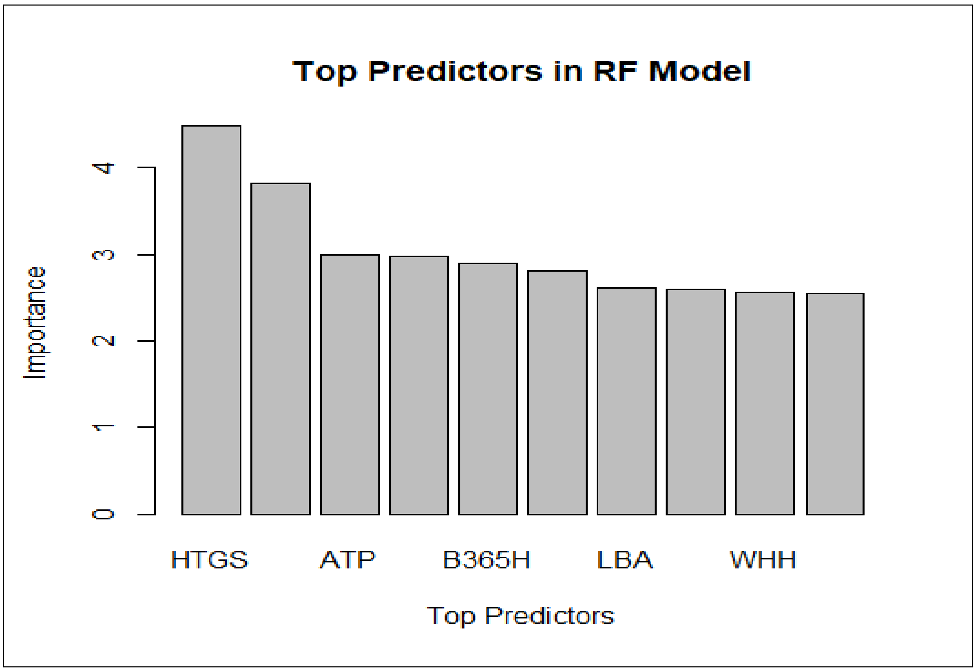


Figure 6: RF AG most important variables

Algorithm	**Target variable**	**Model parameters**	**Performance**
EN	HG	lambda_value <- 0.4 alpha_value <- 0.4	Test R-squared: 0.2069053 Test Mean Absolute Error (MAE): 1.106127
RF	AG	mtry_value <- 20 ntree_value <- 9	Test R-squared: 0.1005539 Test Mean Absolute Error (MAE): 0.9434815

Table 4: End Season Model - Hyperparameters

Strategy	Profit - BB Average (€)	Profit - BB Max (€)	ROI (Average)	ROI (Max)
EN_HG_RF_AG	-61.21	-59.37	-62.21%	-60.37%

Table 5: Performance of the End Season Strategy