UNIVERSITY OF CALIFORNIA

Los Angeles

Goal-Oriented Forecasting: Predicting Soccer Match Outcomes with Deep Learning

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Sheng Chen

2024

ABSTRACT OF THE THESIS

Goal-Oriented Forecasting: Predicting Soccer Match Outcomes with Deep Learning

by

Sheng Chen
Master of Applied Statistics
University of California, Los Angeles, 2024
Professor Guido Montúfar, Chair

Soccer, often referred to as ”football” in the heart of Europe, has a deep-rooted cultural significance that transcends national boundaries. The sport's appeal extends far beyond the pitch, encompassing a wide array of enthusiasts, from die-hard fans to data-driven strategists. In recent years, the fusion of deep learning models with the captivating world of football has taken center stage, revolutionizing our approach to predicting match outcomes. We delve into the fusion of state-of-the-art artificial intelligence and machine learning techniques with the intricacies of a sport that inspires fervent devotion. Our aim is straightforward: to unearth the potential of deep learning models in enriching our capacity to anticipate which team will emerge victorious on the hallowed turf.

We will delineate the primary objectives of this research essay, elaborate on the methodology employed, and elucidate our anticipated contributions to the existing body of knowledge in both the realms of deep learning and sports analytics. Finally, we will underscore the significance of precise football match outcome prediction as an evolving and multi-dimensional research domain that holds great promise for aficionados, professionals, and researchers throughout Europe and beyond.

The thesis of Sheng Chen is approved.

Yingnian Wu

Paik Schoenberg

Guido Montúfar, Committee Chair

University of California, Los Angeles

2024

To my grandpa . . .

who dedicated years fostering

my mathematical education

and patience to overcome any difficulties

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

In the world of soccer, predicting game outcomes and player performances has always been a challenging task. With the increasing availability of data, including statistics, team lineups, and player profiles, there is a growing interest in using data analysis and machine learning techniques to gain insights and make predictions. This document outlines a comprehensive approach to soccer game prediction, focusing on the German leagues from season 2017-2018 to 2021-2022. We aim to understand which team features play the most important role in goal difference, predict future game outcomes, and explore the possibility of predicting games with different starting lineups.

Modern methods are being used to analyze data. Foremost among these is the use of Expected Goals (xG), a statistical tool by Opta that carefully measures how likely a team is to score based on factors including the quality and number of scoring chances created. In the field of betting, where more advanced models have become popular, among which are Machine learning algorithms like Random Forests, Gradient Boosting, and deep neural networks. All these methods share the same idea: use past match data, player performance info, and even factors like weather and injuries to make accurate game predictions. In-play betting has made things more exciting by allowing changes to betting odds while a game is currently underway. Crucial for accurately predicting the outcome, specific measurements are taken to quantify how well both teams and individual players perform. This includes things like where players are on the field, how accurate their passes are, and factors related to their mindset. Furthermore, the industry closely monitors market performance and betting trends, enhancing their ability to comprehend the game's evolution over time.

CHAPTER 2

Prior Work

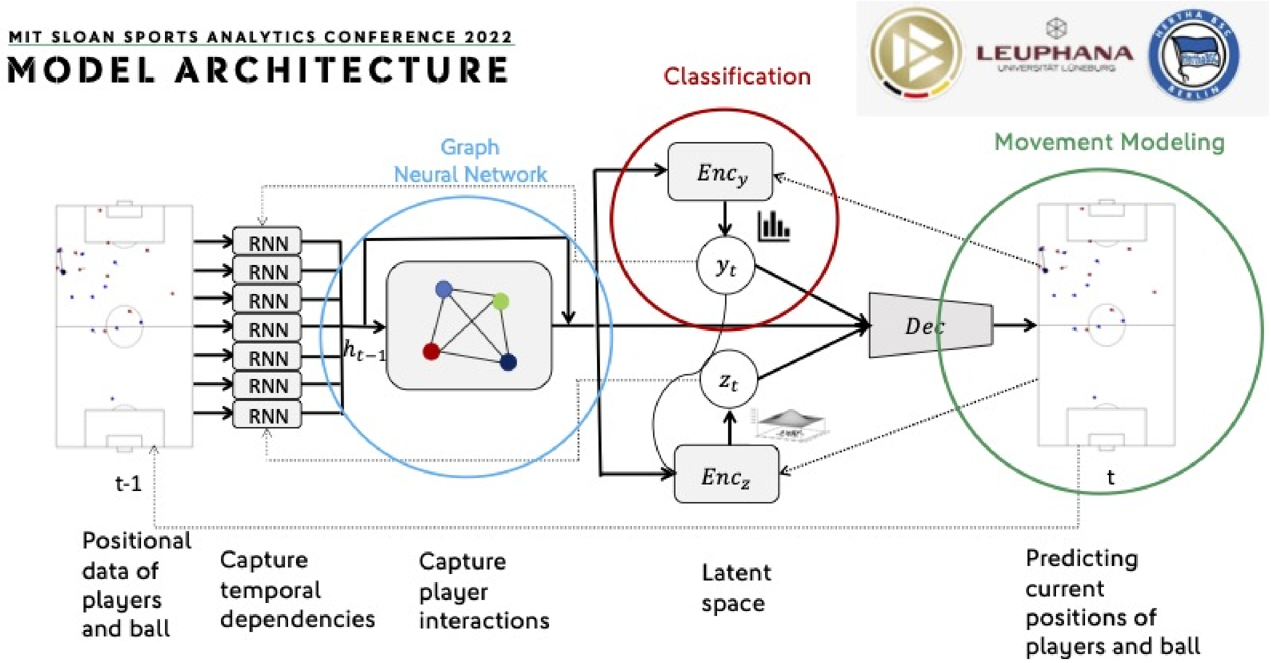
2.1 xG

StatsBomb Expected Goals (xG) is a metric used to assess the likelihood of a soccer shot resulting in a goal. Calculated based on historical data from shots with similar characteristics, an xG model factors in elements such as distance to goal, angle, body part used for the shot, and type of assist. This model assigns a value between 0 and 1 to each shot, representing the probability of it leading to a goal. For a more intuitive understanding, the 2022 update of the StatsBomb xG model gives all penalty kicks a shared static value of 0.78 xG, meaning out of 100 penalty shots, 78 are expected to be scored. According to Lars Maurath's independent research, the alignment between xG and actual goals for team seasons is anticipated to fall within a 95% confidence interval, ranging from 79% to 93%, depending on the quality of the model.

Pascal Bauer and Gabriel Anzer has made progress in the realm of soccer analytics, and notably implemented a Graph Neural Network to automatic detect of tactical patterns with semi-supervised learning. As shown in Fig. 2.1, their work relies on exact positions of players on field as their input, which means they either need high resolution game recordings or player tracker that sends real time location to the data hub. The findings illustrated in Figure 2.1 from the referenced work highlights a dependence on the exact positions of players on the soccer field as crucial input data. This reliance implies a necessity for either high-resolution game recordings or a real-time player tracking system that can relay precise location information to a central data hub. This technological requirement underscores a potential area for future investigation within the domain of soccer analytics. Subsequent research endeavors could explore the feasibility and impact of integrating sophisticated player tracking technologies into predictive models, potentially incorporating advanced deep learning methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and other methodologies. Furthermore, as datasets continue to grow in size and scope over time, the cumulative effect of accumulating substantial data can contribute to the refinement and sophistication of predictive models. The interplay of advanced technologies and the utilization of extensive datasets is promising to the enhancement of the precision and depth of understanding in match predictions. Future researchers may find value in exploring these avenues to further advance the sophistication of analytics in the field. (Bauer & Anzer 2022)

2.2 Video and Image Analysis

In the convergence of computer vision and sports analytics, the combined force of Tracking Data Analysis and Video/Image Analysis emerges as a potent synergy. This integrated approach, propelled by high-resolution videos and computer vision techniques, harmonizes precise player movements and object tracking techniques. By leveraging the strengths of both, these type of research not only advances soccer analytics but also underscores the potential of a unified model in the broader landscape of machine learning-driven computer vision research and real-time sports analysis.



MIT SLOAN SPORTS ANALYTICS CONFERENCE 2022

MODEL ARCHITECTURE

Positional data of players and ball

Capture temporal dependencies

Capture player interactions

Latent space

Predicting current positions of players and ball

Figure 2.1: Detecion of Tactical Patterns using Semi-Supervised Graph Neural Networks (Bauer and Anzer 2022)

CHAPTER 3

Data

3.1 Web Hierarchy

The data utilized in this study was sourced from fbref.com and encompassed statistical information from the five major European football leagues. This dataset covers a span of five seasons, specifically from the 2017-2018 season to the 2021-2022 season. To compile this dataset, information was extracted from a total of 20 league tables. Notably, the dataset comprises detailed information on 380 (or 306 for Bundesliga) matches for each league per year. This comprehensive dataset serves as a fundamental component of the research conducted in this thesis. We employed BeautifulSoup (bs4) to automatically extract links from each league page, which allows us to access detailed information about the games.

3.2 Objective

Which (team) feature plays the most important role in goal difference? Can we predict the goal difference of future games? Can we predict the goal difference of future games given different starting lineups? All three questions can be considered as a self-supervised learning problem since we are using historical data on teams.

Idea for objective 1: This is a classic regression problem with an emphasis on the interpretation of the coefficients, which can give us information on each feature's impact on the output variable.

Idea for objective 2: we can calculate a weighted average of team performance over the past N games to substitute for each row of the model matrix and predict future results.

Idea for objective 3: As an extension of the previous question, we can calculate a weighted average of player performance over the past N games, and this data is used to assess the current form of players in the upcoming games so that based on game dates, we can pull out each player's past five performance and adjust the team data using some function (eg. sum) and adjust the model matrix accordingly. We maintain a "Player Name" dictionary to link players to their respective statistics. A player's level of performance relative to their potential level of performance is called form. Since soccer games are strongly influenced by the current form of the team. Hence, as we consider player stats as a vital aspect of our analysis, we include this factor of fluctuation in performances.

3.3 Organizing Data

We organize data into tabular format, with rows being each game and columns being the team features. To investigate the questions of interest, we need three different matrices. All three matrices should have the same dimensions. The first matrix we need is the team-performance matrix, with rows indicating each game, and columns that include all team features. The columns can also be separated by home and away team features. The second matrix we need is the past-team-performance matrix, with each row replaced by the weighted average of the past n performances from the two teams. N can be a hyperparameter to tune later on. The last matrix we need is the past-player-performance matrix, each row replaced by the aggregate (maybe sum, or other function f) of the weighted average of the past n performances of all players from the two teams. Ideally, the player matrix should give us the most information about the performance and thus predict more accurately, but it is computationally harder to form such a matrix due to data management.

3.4 Exploratory Data Analysis

We investigate the proportion of our output variable, the goal difference, which contains possible values from -9 to +9 excluding -8. If we look at the value counts in Table 3.1, it is

-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
1	0	3	4	44	134	351	762	1481	2229	1815	1072	550	236	99	30	6	4	2

Table 3.1: Outcome with skewed class proportion

easy to notice that there are few cases of +9, -9, and other classes of large absolute values, making it an extremely imbalance classification problem. We have several ways of dealing with such problems, including removing outliers, generating data, or using a weighted loss function.

Examination of specific scatterplots depicting the relationship between the input feature and the output variable of home goals reveals a conspicuous trend. From Fig 3.1, the relationship between of straight corner kicks Teams exhibiting a higher frequency of executing straight corner kicks tend to exhibit lower values in terms of goals. This observation aligns with the judgement that seasoned players possess the proficiency to manipulate the trajectory of the ball, curving it either inward or outward. Such nuanced ball control strategies are evidently superior to the straightforward execution of corner kicks, which, in turn, diminishes the likelihood of scoring. This delineation underscores a noteworthy correlation, indicative of the team's proficiency level, whereas a propensity for straight corner kicks correlates with suboptimal goal-scoring performance, thereby shedding light on specific facets of the game. The analytical framework employed herein substantiates the premise that certain intricacies of the game can be elucidated through systematic analysis, thereby contributing to a more nuanced understanding of soccer dynamics.

3.4.1 PCA

Principle Component Analysis utilizes some of the basic facts in linear algebra to apply data compression. Here, we use PCA to observe the explained variance before we start the modeling. Assume the data consists of points $\{\mathbf{x^{(1)}}, \mathbf{x^{(2)}}, ..., \mathbf{x^{(m)}}\} \in \mathbb{R}^{\mathbf{n}}$. We need some encoding function $\mathbf{f}$ and decoding function $\mathbf{g}$, where $\mathbf{c^{(i)}} = \mathbf{f}(\mathbf{x^{(i)}})$ lies in the lower-dimensional space $\mathbb{R}^{\mathbf{l}}$. For basic PCA, we use the simplest matrix product for the decoder $\mathbf{g}(\mathbf{c^{(i)}}) = \mathbf{D}\mathbf{c^{(i)}}$

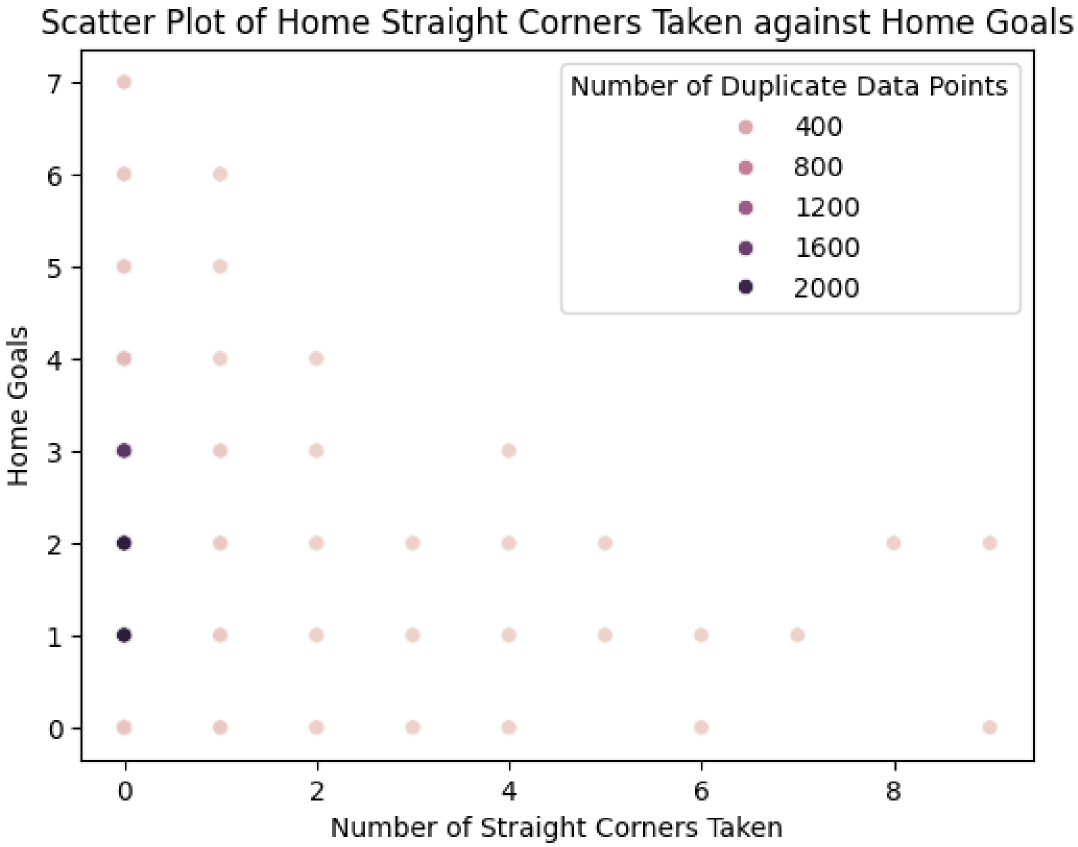


Figure 3.1: Scatterplot of Straight Corners Against Goal Differences

Kernel Function	Accuracy
linear	0.26096866
poly	0.25811966
rbf	0.25754986
sigmoid	0.26267806
cosine	0.25868945

Table 3.2: XGBoost Classfier Accuracy after kernel PCA

with $\mathbf{D} \in \mathbb{R}^{m \times l}$. The goal is to minimize the construction loss:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \|\mathbf{x} - \mathbf{g}(\mathbf{c})\|_{\mathbf{2}}^{\mathbf{2}}$$

where we find the By solving the above, we get $\mathbf{c} = \mathbf{D^T x}$, then reconstruction loss becomes

$$\arg\min_{D} \|\mathbf{X} - \mathbf{D D^T X}\|_{Frobenius}$$

, where $\mathbf{X} = [\mathbf{x^{(1)}}\, \mathbf{x^{(2)}} \ldots \mathbf{x^{(m)}}]^{\mathbf{T}}$. This optimization problem is solved through eigendecomposition with the optimal D obtained through vertically combining the n eigenvectors of $\mathbf{X^T X}$ that corresponds to the n largest eigenvalue.

In kernel PCA, we replace $\mathbf{X}$ with function $\phi(\mathbf{X})$ that is able to operate in high dimensions, without ever computing in that space, but rather the dot product in a low dimension. If we use PCA with different kernel functions, we get the following logistic regression accuracy in Table 3.2 .

Despite the application of diverse kernel transformations in the Principal Component Analysis (PCA) process, the restricted use of 5 principal components for computational efficiency yields unsatisfactory results, even when employing the best-performing model, which, as will be detailed subsequently, is XGBoost. This indicates that the limited dimensionality represented by 5 principal components falls short in adequately encapsulating the richness of information within the data distribution. In the intricate domain of soccer, where numerous features equally impact the game, each action contributing uniquely to the outcome, it becomes apparent that a higher-dimensional representation is imperative for a more comprehensive understanding of the underlying dynamics. The inadequacy of 5 principal components underscores the necessity for a more expansive representation, emphasizing the need for additional dimensions to capture the nuanced relationships within the dataset.

CHAPTER 4

Methodology

Given our treatment of the outcome variable, goal differences, as a discrete variable, the task at hand transforms into a classification problem. Upon obtaining our training data, the initial imperative lies in addressing the challenge posed by imbalanced classification. Traditionally, there exist two primary approaches to tackle class imbalance: the first involves employing data augmentation techniques, while the second entails utilizing loss functions with re-weighting terms. While some resampling techniques offer advantages, it is essential to acknowledge their inherent limitations (there is no free lunch). The most basic approach to over-sampling involves duplicating random records from the minority class, potentially leading to overfitting. Conversely, in under-sampling, the simplest technique entails removing random records from the majority class, a practice that may result in a loss of valuable information. Here, we choose to apply several mixture of data augmentation techniques, and later we compare the results on modeling testing data with the method without augmentation but with a re-weighting loss function.

4.1 Data Reduction

4.1.1 Tomek links

We start by implementing Tomek links, a simple undersampling technique, referring to pairs of instances that are in close proximity but belong to opposing classes. By eliminating instances from the majority class within each pair, the separation between the two classes is enhanced, thereby streamlining the classification process. A Tomek link is identified when two samples emerge as the nearest neighbors of one another. In leveraging this method for undersampling, the objective is to strategically reduce the dominance of the majority class, fostering a more balanced and discriminative dataset for classification tasks.

4.2 Data Augmentation

To apply some of the techniques below, there is a minimum number of points required (theoretically we can do augmentation even with only 2 points for a given class, but with such a lower number of observations, newly generated points would all be near the line connecting the original two points). For our dataset, which contains some classes with numbers of observations less than 10, it is reasonable to remove these extreme outliers in our classification modeling and left us with the outcome distribution shown in Fig.4.1 .

4.2.1 SMOTE and its variations(borderline, ADASYN)

The Synthetic Minority Over-sampling Technique(SMOTE) is a technique to generate synthetic tabular data. All variations of this oversampling technique create points by using a convex combination of points that belong to the selected minority classes, and by repeatedly applying this method to expand all classes to the same number of counts, which, in our case, is 2229.

Quickly running a linear regression after borderline SMOTE gives us fig. 4.2

4.2.2 Tabular GAN

For tabular GAN and Conditional GAN, both methods emulate the distribution of the training matrix and outcome vector. Again, we are attempting to generate data using TGAN so that each outcome class has an equal count. Therefore, we split the data one by one for each class and generated the required number for each class.

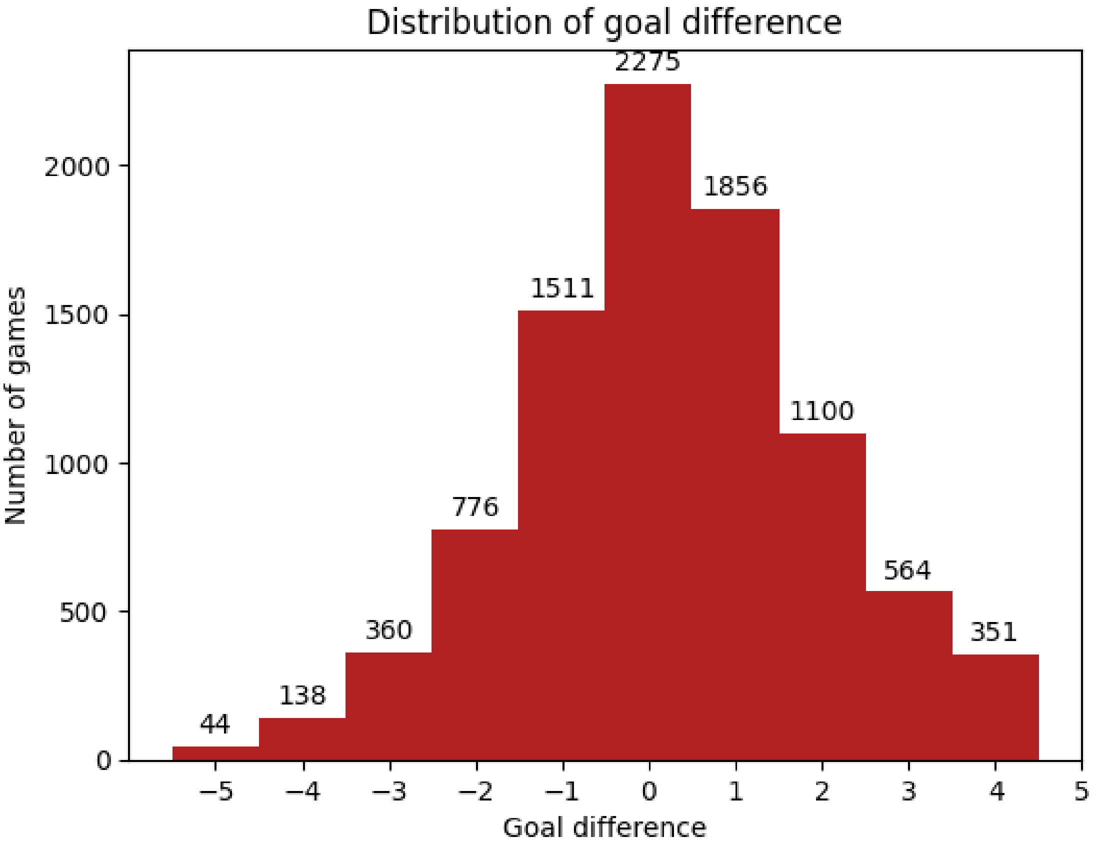


Figure 4.1: Histogram of Goal Differences

4.3 Feature Selection

In our case, feature selection is used as a preprocessing step in conjunction with machine learning models for regression/classification purposes. The purpose is to find the best feature subset with a determined number of features. Feature selection methods can be classified into wrapper, embedded, filter methods, or the hybrid of the previous three methods. Wrapper methods treat the classifier as a black box and select features based on the classification result using some metric (eg. forward, backward selection); embedded methods carry out selection through the classifier algorithm itself (e.g. regularization, tree-based algorithms); filter methods have the selection process independent of classifier algorithms (e.g. univariate ANOVA). [feature selection review] Typically, filter methods are the fastest, and wrapper the slowest.

4.3.1 Exhaustive Search

In general, exhaustive search over feature space is often impractical due to combinatorial explosion. The feature space refers to all the subset of the features, and the number of such combinations grows exponentially with the number of features. For example, with 20 features, there are 2^{20} possible subsets to consider. The computational cost increases exponentially with each additional feature, making an exhaustive search over the feature space quickly unmanageable. Some feasible approaches to get around is to assume a fixed number of features to select, denote it as k. We can implement a greedy search of $\binom{n}{k}$ number of possible subsets of size k to find the optimal subset based on the metrics we choose.

The combinatorial explosion is a fundamental challenge in feature selection, and it highlights the need for efficient and heuristic methods to identify relevant features without exploring the entire feature space. Techniques such as greedy algorithms, recursive ffeature elimination, or optimization methods are often employed to strike a balance between computational complexity and obtaining a good subset of features for predictive modeling. In this case, to implement a simple exhaustive search, we set the subset size fixed as k = 5 for the sake of computational efficiency.

4.3.2 Lasso

Lasso, or L1 regression, is a common regularization technique based on linear regression. It adds a weighted L1 norm of the coefficient vector to the least squares function. It is not hard to notice, once the weight was set large enough, some elements in the coefficient vector will diminish to zero, and thus performing feature selection while doing regression(embedded feature selection). However, the downside of using lasso is the its lack of stability and its natural assumption of a linear relationship, which stems from its fundamental nature of being a linear regression .

4.3.3 Mutual Information

Mutual information (MI) is a measure of the amount of information that one random variable has about another variable, and in our case, it is used to quantify and compare the relevance of different feature subsets concerning the output variable. It is defined to be

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \,||\, P_X \otimes P_Y) \tag{4.1}$$

where $(X,\, Y)$ are pairs from the space $\mathcal{X} \times \mathcal{Y}$, $P_{(X,Y)}$ is the joint probability mass function, and $\otimes$ indicates outer product, and in univariate case, simply the scalar product of the marginal distribution $P_X(x) \times P_Y(y)$.

The Kullback-Leibler (KL) divergence measures the difference between two probability distributions. For discrete probability distributions P and Q, the KL divergence ($D_{\mathrm{KL}}(P \,||\, Q)$) is defined as:

$$D_{\mathrm{KL}}(P \,||\, Q) = \int_{\mathcal{X}} \int_{\mathcal{Y}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx\, dy \tag{4.2}$$

, and for continuous variable, integrals instead of summations.

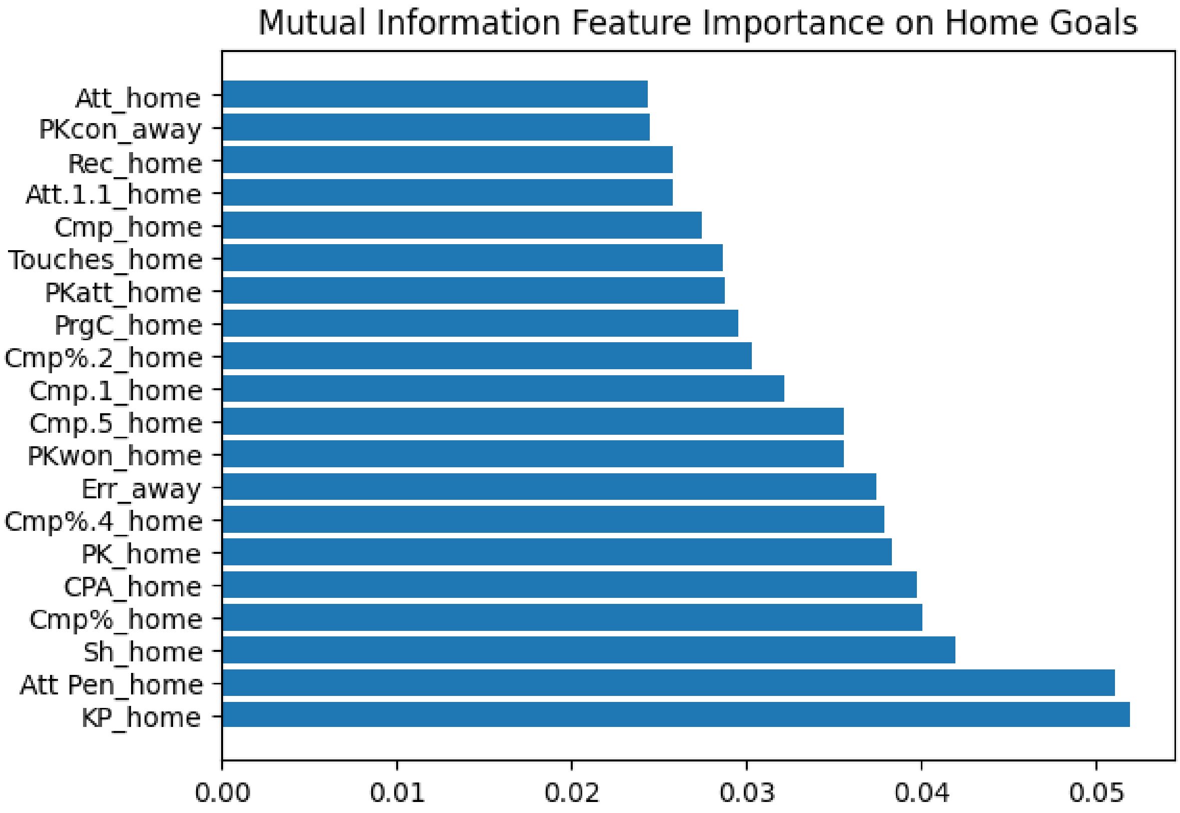


Figure 4.2: Bar Plot of Mutual Information of Input Features on Home Goals

Thus, mutual information can be written explicitly:

$$I(X;Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{(X,Y)}(x,y) \log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right) dx\, dy \tag{4.3}$$

where $P(X,Y)$ is the joint probability distribution of X and Y, and $P(X)$ and $P(Y)$ are the marginal distributions.

Since all of our dataset's features are continuous, we need nonparametric methods based on entropy estimation from k-nearest neighbors distances as described in Estimating Mutual Information and mi between discrete and continuous. The result is implemented by the mutual_info_classif function in scikitlearn.feature_selection. We apply Mutual Information between each feature and output variable to perform this filter method.

In order to empirically elucidate the significance of mutual information among input features concerning home and away goals, our study adopts home goals as the output variable.

As delineated in Figure 4.2, while the mutual information (MI) values across features do not exhibit conspicuous prominence, an analysis reveals that pivotal features, notably attacking maneuvers from the home side encompassing key passes, touches in the penalty area, and total shots, alongside errors committed by opponents and penalties awarded to the away team, exert a discernible influence on home goal outcomes. The observed results align with our a priori expectations, thereby reinforcing our intuitive understanding. Furthermore, our findings establish a robust statistical foundation elucidating the salience of these factors in influencing home goals. This not only enhances our comprehension but also furnishes statistical validation for the pivotal role played by these factors. Subsequently, scouts or analysts may leverage these metrics to make informed decisions regarding the identification of key players on the field or determining the corresponding player's salary based on their contribution to the team's results from a statistical standpoint.

In the context of implementing feature selection using mutual information, the approach to selecting features, whether sequentially adding them or considering mutual information with respect to the target variable Y in one go, can yield significant differences in outcomes. Whether or not to augment the data before feature selection can also lead to a different result, as shown in Fig. 4.3 .

When features are sequentially added, similar to the adapted method shown in greedy search in section 4.3.1, the selection process involves iteratively adding features based on their individual mutual information with Y. This method allows for a step-by-step refinement of feature selection, where each added feature is chosen based on its incremental contribution to mutual information. On the other hand, considering mutual information with Y in one go involves evaluating the mutual information for all features collectively. This holistic approach considers the combined information provided by the entire set of features in relation to the target variable. To illustrate this process, we generate a barplot where each bar represents the information gain contributed by an added feature. This step-by-step addition of features results in a non-decreasing barplot, showcasing the cumulative improvement in information gain as we progressively incorporate more features into the model. This approach provides

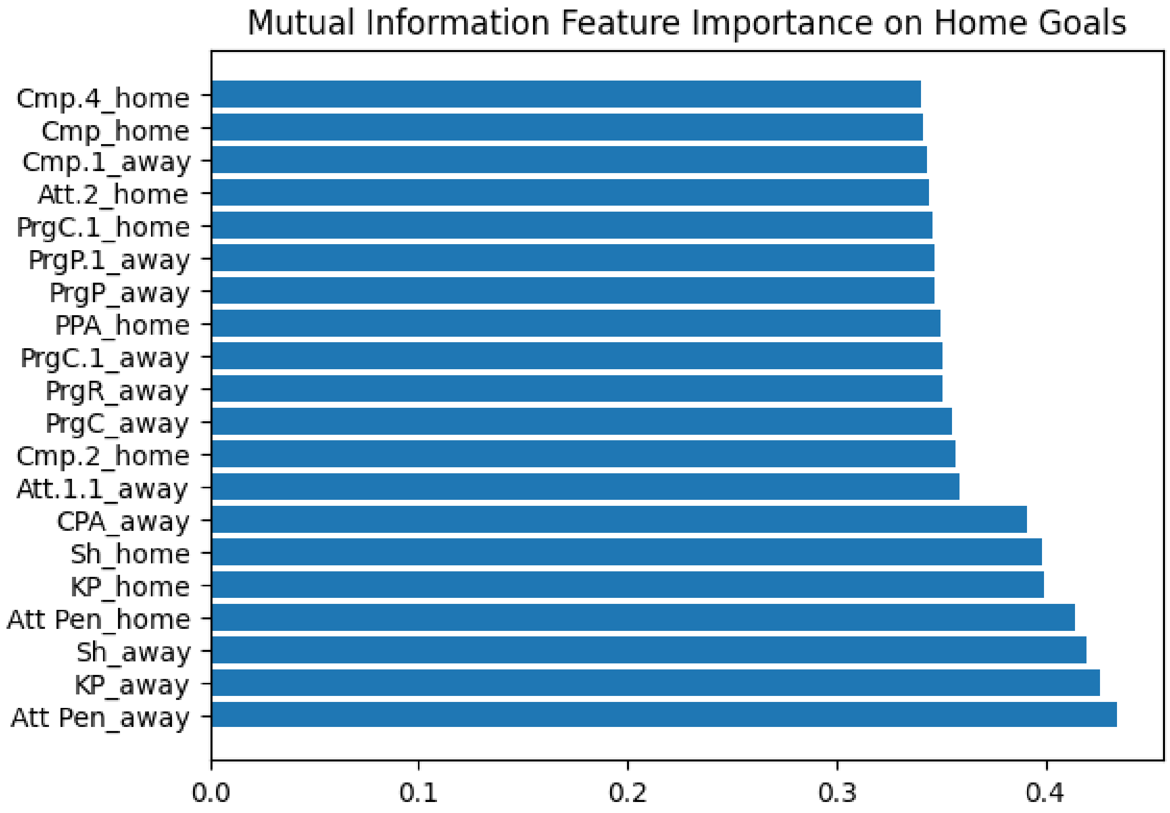


Figure 4.3: Bar Plot of MI gain of added Input Features on Home Goals

a clear visual representation of the utility of each feature, allowing us to discern trends and identify the point at which additional features contribute less substantially.

As we observe the barplot, we eventually reach a saturation point, where the incremental information gain diminishes, indicating that further addition of features does not significantly improve predictive performance. At this juncture, we strategically decide on a threshold, beyond which we cease adding features. This threshold is a crucial determinant in streamlining the feature set for optimal performance without unnecessary complexity. The chosen metrics for evaluating the effectiveness of this feature selection process is the direct result from XGBoost (XGB) modeling over possible feature set.

The choice between the above two strategies can impact the set of selected features and, consequently, the model's performance. A sequential addition may capture nuanced relationships incrementally, while a simultaneous evaluation may reveal synergies or dependencies among features that contribute collectively to mutual information with Y.

4.4 Models

4.4.1 Linear Regression

Linear regression is the fundamental statistical method used to model the linear relationship between an outcome variable with independent variables. Geometrically speaking, the model helps us find the best fitted line that explains the data points in a way that minimizes the distance between the predicted and the actual data points. The equation of the line is commonly expressed as $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{y}$ is the outcome variable, $\mathbf{X}$ is the model matrix, β is the slope of the line, and ϵ denotes the random noise.

4.4.2 XGBoost

Extreme Gradient Boosting, or XGBoost, is the state-of-the-art ensemble tree algorithm. It utilizes the concept where weak learners (in this case usually simple decision trees) are used sequentially to construct a strong learner. XGBoost enhances this process by incorporating ridge and lasso regularization techniques, enables parallel and distributed computing, and combines Newton's method and gradident descent to accelerate optimization speed. This method is known for its exceptional speed and ability to handle large datasets on tabular data for both regression and classification tasks.

Softmax Objective Function:

$$\text{Objective} = -\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_{ik}\log(p_{ik}) + (1-y_{ik})\log(1-p_{ik})\right] + \sum_{k=1}^{K}\Omega(f_k) \tag{4.4}$$

4.4.3 MLP

Multilayer Perceptron (MLP), involves stacking layers of linear regression with activation function in order for it to learn patterns from data. While its basic structure is straightforward, the model's simplicity conceals its capability to capture complex relationships in diverse datasets, making it a foundational and accessible entry point into deep learning.

4.4.4 Tabular Transformer

TabTransformer (Huang et al. 2020), utilizes self-attention based Transformers to enhance predictive accuracy on tabular data. The architecture includes a column embedding layer, a series of Transformer layers transforming categorical features into contextual embeddings, and a multi-layer perceptron (MLP). The concatenated contextual embeddings and continuous features are fed into the MLP, and the model is trained end-to-end by minimizing the loss function to learn all parameters.

CHAPTER 5

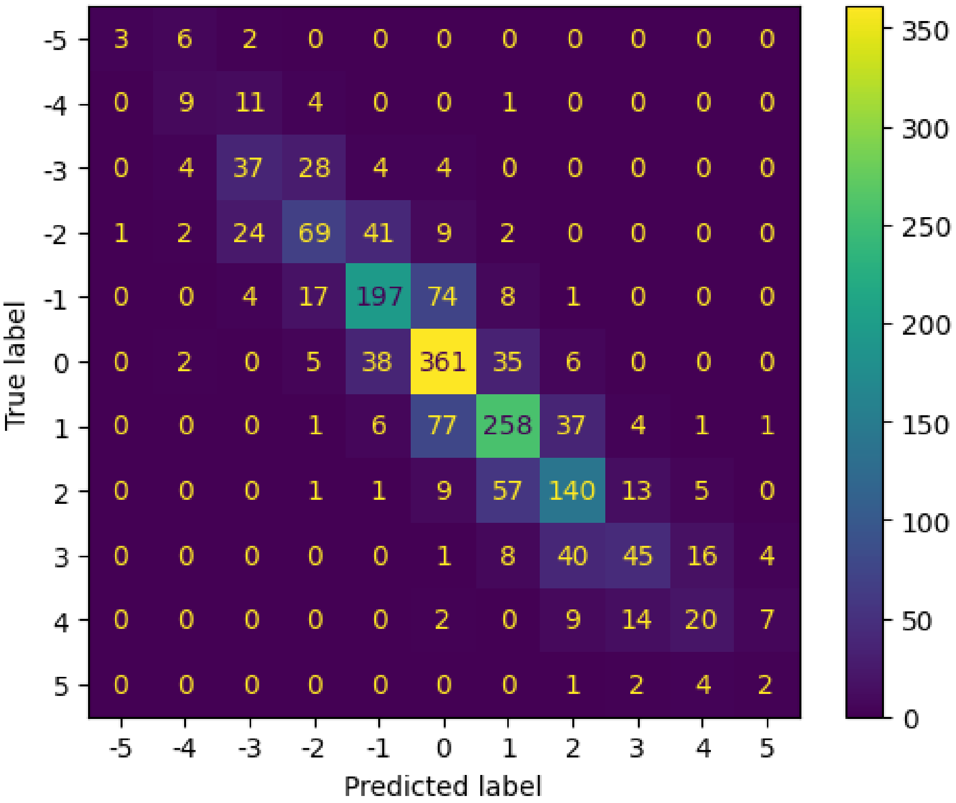
Result

In our comprehensive experimentation, we investigate various combinations of data augmentation techniques, feature selection methods, machine learning algorithms, and loss functions to evaluate their impact on predictive modeling. The primary metric for comparison is the accuracy score, and our objective is to discern optimal configurations for improved performance.

A crucial consideration in our experimental design is the order of applying data augmentation and feature selection. We recognize that if feature selection precedes data augmentation, there is a risk of introducing bias in the generated data. Augmenting the data before feature selection, however, allows us to capture more information, potentially enhancing the richness of the dataset and leading to more informed model selection. Table 5.1 encapsulating key performance metrics for each experimental configuration.

Step1	Step2	Method	Loss Function	Accuaracy
SMOTE	MI	Linear Regression(Round)	Ordinary Least Squares	0.1267
SMOTE	MI	Logistic Regression	Cross-Entropy	0.2093
SMOTE	MI	XGB	Softmax	0.3450
/	MI	Linear Regression	Weighted Least Squares	0.3680
/	MI	MLP	Weighted Least Squares	0.2373
MI	SMOTE	XGB	Softmax	0.4344

Table 5.1: Model Result with Combination of Methods



	-5	-4	-3	-2	-1	0	1	2	3	4	5
-5	3	6	2	0	0	0	0	0	0	0	0
-4	0	9	11	4	0	0	1	0	0	0	0
-3	0	4	37	28	4	4	0	0	0	0	0
-2	1	2	24	69	41	9	2	0	0	0	0
-1	0	0	4	17	197	74	8	1	0	0	0
0	0	2	0	5	38	361	35	6	0	0	0
1	0	0	0	1	6	77	258	37	4	1	1
2	0	0	0	1	1	9	57	140	13	5	0
3	0	0	0	0	0	1	8	40	45	16	4
4	0	0	0	0	0	2	0	9	14	20	7
5	0	0	0	0	0	0	0	1	2	4	2

Figure 5.1: XGBoost Confusion Matrix with accuracy 0.63

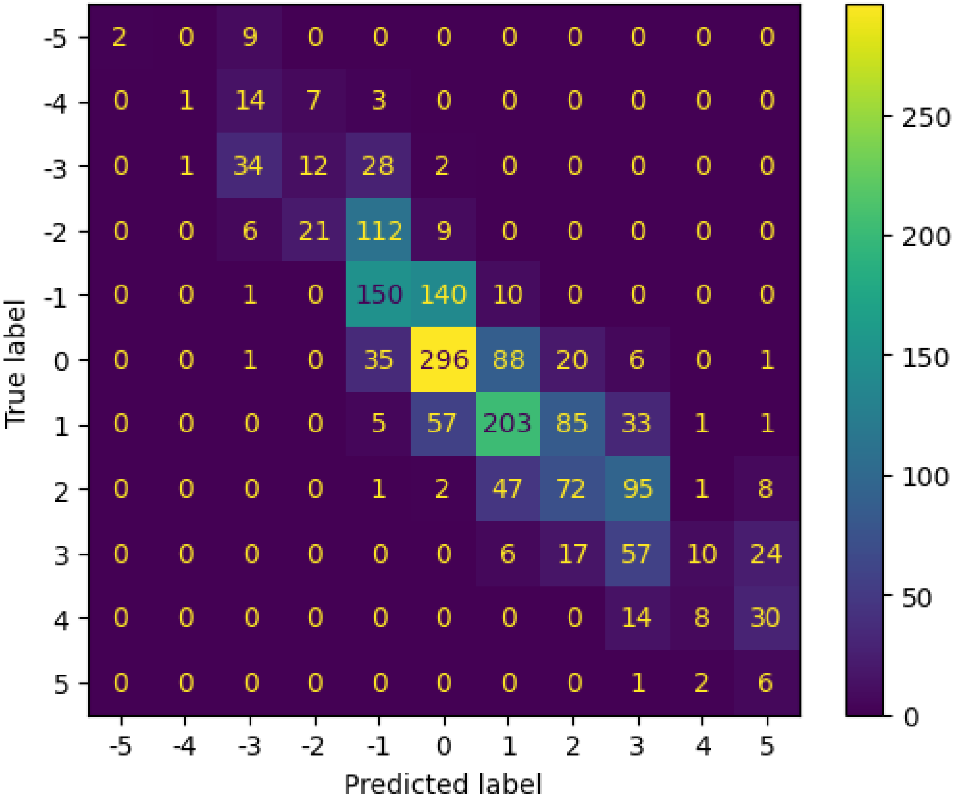


Figure 5.2: MLP Confusion Matrix with accuracy 0.474

5.1 MLP

hidden_layer_sizes=(100,70,100), max_iter=1000

Accuracy: 0.474

CHAPTER 6

Future Work

The study strategically fused cutting-edge artificial intelligence and machine learning techniques with the intricacies of soccer, focusing on the prediction of match outcomes. As a potential direction for future researchers, the incorporation of real-time data streams during matches is suggested, enabling dynamic model adjustments based on evolving game scenarios. With technological advancements, the increased computational power afforded by robust GPUs enables the exloration of a myriad of model architectures. This empowerment allows researchers to investigate not only the depth and complexity of CNNs but also the sequential and temporal dependencies captured by RNNs and the diverse capabilities of other advanced models like Transformer networks. Furthermore, the synergy between powerful GPUs and the surge in the volume and precision of data collection across multiple avenues is pivotal. The amalgamation of meticulously curated datasets, spanning player movements, team statistics, and contextual factors, establishes a foundation for model training characterized by richness and diversity. This wealth of high-quality data not only reinforces the resilience of the models but also facilitates the exploration of intricate relationships within the multifaceted realm of soccer. Advancements in data collection methodologies, including real-time player tracking systems and high-resolution game recordings, further elevate the potential for refining deep learning models. The precision in data acquisition empowers researchers to capture detailed aspects of player dynamics, fostering a more comprehensive and accurate representation of the intricate dynamics inherent in soccer matches.

Another area ripe for improvement involves the exploration of ensemble models, combining predictions from diverse algorithms to enhance overall accuracy. To enhance interpretability, methodologies like SHAP values are recommended to illuminate the significance of different features in decision-making. The potential integration of human-in-the-loop approaches, involving expert insights in model training, presents an avenue for refinement. Transfer learning, the creation of user-friendly interfaces for stakeholders, and ethical considerations in predictive model applications are also critical areas for future exploration. By delving into these directions, researchers can contribute to the ongoing evolution of goal-oriented forecasting in soccer analytics.

REFERENCES

[ABB22] Gabriel Anzer, Pascal Bauer, Ulf Brefeld, and Dennis Faßmeyer. "Detection of tactical patterns using semi-supervised graph neural networks." 03 2022.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[KL87] L F Kozachenko and N N Leonenko. "Sample Estimate of the Entropy of a Random Vector"." *Probl. Peredachi Inf*, **23**(2):95–101, 1987.

[KSG04] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information." *Phys. Rev. E*, **69**:066138, Jun 2004.

[Ros14] Brian Ross. "Mutual Information between Discrete and Continuous Data Sets." *PloS one*, **9**:e87357, 02 2014.

[XS21] Peter Xenopoulos and Claudio Silva. "Graph Neural Networks to Predict Sports Outcomes." In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1757–1763, 2021.