Visual Analytics in Sports Betting: Analyzing and Comparing Odds using Dimensionality Reduction

Giovanni Pignata
Visual Analytics 23/24
Engineering in Computer Science
University of Rome "La Sapienza"

February 2024

Abstract

This study presents an innovative web application designed to apply visual analytics and dimensionality reduction techniques, specifically utilizing D3.js and Druid.js, to the analysis of sports betting odds. Through interactive data visualization, the application allows users to identify clusters within betting data, revealing odds with potentially high or low value relative to the actual probabilities of the outcomes. The results highlight the power of visual analytics in discovering advantageous betting opportunities, thereby assisting users in making informed decisions. The source code is available on https://github.com/Zigarov/Odds-Visual-Analytics.

Keywords: visual analytics, betting odds analysis, sports betting, dimensionality reduction, D3.js, Druid.js.

1 Introduction

The sports betting market has witnessed substantial economic growth, highlighting the increasing engagement of individuals in sports betting activities. This study is aimed at bettors who want to move away from impulsive betting and adopt a more analytical approach based on data analysis. This transition underscores the evolving nature of sports betting from a mere game of chance to a data-driven decision-making process.

Despite the proliferation of studies and tools aimed at statistically analyzing matches to enhance prediction accuracy, there remains a notable gap in the literature concerning the analysis of betting odds. While tools to dissect match strategies and outcomes abound, resources dedicated to scrutinizing and comparing the distribution and intrinsic values of different types of betting odds are conspicuously scarce. This oversight is significant, given that the value of odds does not always reflect the actual probability of the event occurring, thus impacting the expected return on the long run.

The goal of this study was to develop a suite of visual analytical tools enabling bettors to compare odds across various match types and raise the level of their decision making by exploiting possible patterns within betting odds distributions. The proposed tools offer a novel approach to facilitates both outlier identification and cluster analysis by employing a dimensionality reduction algorithm. Thus allowing users to visually compare odds distribution, the derived probability of specific outcomes and their effective recurrences whitin the dataset.

The application of the developed tools to analyze the principal odds (1,x,2 over2.5, under2.5) for matches in the 2022-2023 season across the top five national leagues (Italy, England, Spain, Germany, France), has made it possible to identify various clusters of events that offer a higher expected return compared to others, showcasing the practical utility of the developed methodologies in real-world betting scenarios.

By addressing a previously under-explored aspect of sports betting analytics, this study aims to make a significant contribution to the field, offering a new perspective on the comparative analysis of betting odds

2 Background

This section explores the state of the art regarding Visual Analytics in Sports Betting. In addition, technical solutions similar to those implemented in this study are presented.

2.1 Visual Analytics in Sports Betting

Recent surveys in the domain of sports analytics, particularly those focused on football, have shed light on the newest developments in data analysis, visualization, and predictive analytics [6, 1]. These comprehensive reviews highlight the growing role of visual analytics in transforming raw data into

actionable knowledge for decision-making tasks like betting. Despite the broad array of methods and tools proposed, it is observed that the state of the art focus predominantly on the analysis of sports performance data rather than exploring the betting odds provided by bookmakers. This distinction points to an under-explored area within the realm of sports betting analytics, where the potential for visual analytics to unlock new insights into betting strategies remains largely unexploited.

2.2 Related Technical Solutions

Parallel Coordinates Plot The Parallel Coordinates Plot (PCP) is a visualization technique that represents multivariate data as lines across parallel axes, each corresponding to a dimension of the dataset. This method is particularly revered in the field of visual analytics for its ability to reveal patterns, relationships, and outliers in high-dimensional data. The design of the PCP in this project draws inspiration from the **High order brushing** technique proposed by [4]. High order brushing enhances the interactivity of PCPs by allowing users to select and highlight data subsets based on complex, multi-dimensional criteria. This advanced form of interaction extends the utility of PCPs beyond simple visualization, enabling a more nuanced exploration and analysis of the data.

Dimensionality Reduction Dimensionality Reduction (DR) is a core building block in visualizing multidimensional data. Finding the similarities and differences between groups of datasets is a fundamental analysis task. For this reason, many visual analytics systems have already demonstrated the benefits of tightly integrating DR with interactive visualizations as widely reviewed in [5]. This project implements Druidjs [3], an open-source library developed in JavaScript, capable of computing dimensionality reduction directly in the browser with performance comparable to other libraries such as Scikit-learn [3, 2]. Although the projections in the experiments were pre-computed to conserve resources, the adoption of Druid.js is motivated by the future prospect of allowing users to upload their custom datasets. This flexibility paves the way for leveraging client-side JavaScript's interactivity, offering users a personalized and dynamic visual analytics experience.

Probabilities Derived From Betting Odds A study conducted in 2016 introduced various approach for analyzing probabilities derived from betting odds and their associated biases [7]. Among these methods, one aligns with the approach utilized in the current study, reinforcing the relevance of this

analytical technique. However, the paper also outlines alternative methods that could potentially offer superior performance. This revelation opens up future investigation.

3 Methodology

Traditional approaches to betting, usually lead by intuition, often neglect the study of odds and their intrinsic value in relation to the probability of that particular outcome occurring. Thus, the primary objective of this study is to pivot the betting strategy from a guesswork-based approach to a data-driven methodology, enabling users to make informed decisions by analyzing and comparing the distribution of odds offered by bookmakers.

The architecture of the web application is designed to be user-friendly while harnessing robust analytical tools that provide deep insights into betting odds. It is structured into two main components: the back-end and the front-end.

Back-end The back-end, serving as the application's backbone, is responsible for data management, processing, and analysis. Developed using Node.js and employing Web-pack for module bundling, it facilitates efficient handling of large datasets from various bookmakers and leagues. The back-end also incorporates the use of Druid.js for implementing the **Multidimensional Scaling algorithm** (MDS), transforming multidimensional odds data into a visually interpretable form. This computational layer prepare the data for interactive exploration on the front-end.

Front-end The front-end is the interface where users interact with the application, visualizing the processed data through dynamic and interactive components. The interface is composed by three interactive and coordinated visualizations, entirely developed using D3js, a powerful tool for creating complex and responsive data visualizations. The integration of these visualizations allows to engage with the data in an intuitive manner.

3.1 Visualizations

The interface is composed by three interactive and coordinated visualizations:

- **Parallel Coordinates Plot**: Offers an overview of the dataset, allowing users to grasp the multidimensional nature of the data at a glance.

- **Scatter Plot with Dimensionality Reduction**: Visualizes the data in 2 dimensions after the projection with the MDS algorithm, enabling users to identify clusters, patterns, and outliers in the odds distribution.
- **Comparative Chart**: Provides a comparative analysis of odds distribution for different match outcomes through box-plots, revealing the derived probability distribution and highlighting key statistical metrics.

These components are designed to be interactive and coordinated, allowing users to select specific subsets of data for detailed analysis.

3.1.1 Parallel Coordinates Plot

The Parallel Coordinates Plot (PCP) is a common method for visualizing high-dimensional data by drawing each data point as a line that intersects multiple parallel axes, each representing a dimension of the data. This visualization technique is particularly useful for identifying patterns, correlations, and potential outliers across multiple dimensions, facilitating a comprehensive analysis of complex datasets.

Design The implementation of the Parallel Coordinates Plot for this web application was possible thanks to D3.js. The plot was designed to display five vertical and parallel axes, corresponding to the dimensions most relevant to the analysis, the betting odds for the available outocome: AvgH represents the average of the odds for the home team win, AvgD for the draw, AvgA for the away team win, AvgO represents the average of the odds for the total number of goals greather than 2, and AvgU represents the contrary.

Interactions As a key interactive and coordinated feature, the application allows users to select desired value ranges on each axis through a multi-dimensional brushing functionality. Brushing in a Parallel Coordinates Plot enables users to specify intervals on the axes by clicking and dragging across the axis length, effectively filtering the data displayed. Lines that meet the selection criteria are highlighted in both the Parallel Coordinates Plot and the Scatter Plot. The corresponding data rows are then used for further analytics, which are visualized in the Comparative Chart. In addiction, the user can use the focus function in order to filter the data used to plot the PCP to only the ones selected by the brushing, allowing a deeper exploration inside the data distribution. This interactive brushing capability significantly enhances the user's ability to explore and analyze the dataset, enabling a dynamic and user-driven approach to uncovering insights within the sports betting data.

3.1.2 Scatter Plot

A Scatter Plot is a type of data visualization that uses dots to represent values obtained from two different variables, allowing the observer to detect any patterns, trends, or correlations between them. For this project, the scatter plot visualizes points in a two-dimensional plane, with each point's position derived from the application of a dimensionality reduction algorithm, specifically Multidimensional Scaling (MDS), to the same columns of the dataset used in the parallel coordinates plot.

Dimensionality Reduction and MDS In the field of visual analytics, dimensionality reduction algorithms are very useful for simplifying complex, high-dimensional data by mapping them into a lower-dimensional space that can be easily analyzed and visualized. These algorithms facilitate the identification of patterns, trends, and relationships that might not be visible in the original multidimensional data, making them crucial for data exploration and decision-making processes. Multidimensional Scaling (MDS) is a statistical technique designed for analyzing similarity or dissimilarity data, aiming to position each item in a lower-dimensional space while preserving the pairwise distances between items, thus enabling a faithful representation of the dataset in a two-dimensional space. This characteristic makes MDS one of the most appropriate algorithms for the purpose of the study, facilitating the exploration and understanding of multidimensional betting odds data. The core of MDS revolves around the distance matrix, as it encapsulates the entire dataset's structure in terms of the distances between items, according to a specific metric, in this case the simple Euclidean Metric. This matrix is then used by MDS to project the items into a lower-dimensional space (typically two-dimensional for visualization purposes) in a way that attempts to preserve these pairwise distances. In the implementation, data normalization through Z-score was a crucial pre-processing step to ensure accurate visualization and prevent any dimension from dominating others. Then, Druid.js was used to compute the distance matrix and generate the two-dimensional projection. Druid.js is a JavaScript library known for its efficiency in computing advanced mathematical operations, making it an ideal choice for handling the computational complexity of MDS. By applying MDS with the Euclidean distance metric, complex, high-dimensional betting odds data are transformed into a simplified, yet insightful, two-dimensional scatter plot.

Visual Encoding The design of the scatter plot was carefully considered to provide insightful visual cues about the match outcomes. Points are color-

coded to indicate the winner of the match (red for the home team, green for the away team, and white for a draw), while the shape of the point reveals whether the number of goals scored was over or under the 2.5 threshold. Specifically, matches with more than 2 goals are represented as circle, and those with 2 or fewer goals as triangles. This encoding allows users to quickly grasp the nature of each match outcome at a glance.

Interactions The scatter plot incorporates bi-dimensional brushing, enabling users to select specific data points by drawing a rectangle over the desired area. This selection not only highlights the points within the boundary in the scatter plot but also emphasizes the corresponding lines in the parallel plot. The selected data are further utilized for analytics, which are subsequently displayed in the comparative chart, ensuring a cohesive and informative exploration. The focus function is implemented also for the bi-dimensional brushing, allowing users to better explore the data selected in the brushing area inside the parallel coordinates plot.

3.2 Comparative Chart

The objective of the Comparative Chart within this web application is to equip users with graphical tools for visually analyzing the distribution of the selected betting odds. The chosen instrument for this analysis is the box-plot, a standardized way of displaying the distribution of data. Box-plots are particularly useful for showing the range and distribution of data points across different outcomes (H, D, A, Over, Under), making it easier for users to assess the variability and central tendency of betting odds.

Probability Distribution from Betting Odds Instead of directly considering the distribution of betting odds to calculate the necessary metrics for drawing the boxplot, this work opts to utilize the derived success probability defined as $p_q = \frac{1}{o}$, interpretable as the hypothetical probability for which the outcome associated with betting odds o has non negative expected return value. This approach aims to analyze the intrinsic value that an odds represents. The induced success probability by o is directly comparable to the frequency of the associated outcome in the considered data distribution, computed as the ratio of the number of matches that resulted in that outcome over the total size of the selected dataset. An additional advantage of inverting the original values of the distribution is that it standardizes the value range across all considered dimensions, facilitating visual comparison between different outcomes.

Analytics For each dimension, two boxplots are drawn. In the firts box-plot, the first and third quartiles metrics define the box's length and position, while the median is represented by a continuous line cutting across the box-plot's width. In the second boxplot, drawn with a dashed line and half the width with respect to the first one, the standard deviation defines the box's height and position on the axis. The distribution's mean is represented as a dashed line cutting across the second box-plot. Whiskers are calculated as the maximum and minimum values within the inter-quartile range. Values outside this range are considered outliers and represented as points on the specified axis. Additionally, a red cross is drawn on the axis, at a position corresponding to the actual frequency of that outcome occurring within the considered distribution.

User Interactions Besides selecting which metrics to display in the box-plot (median and quartiles, mean and standard deviation, or both), an on-MouseOver event was developed for each drawn box-plot. This feature ensures that when the cursor hovers over the drawn area, a tool-tip appears on the screen displaying the values of each calculated metric, other the cardinality of the selected distribution. This interactivity enhances the user experience by providing immediate access to detailed statistical information, further aiding in the comprehensive analysis of the betting odds' distribution.

4 Experimental Result

To test the functionality offered by the designed tools, experiments were carried out using data from the 2022-2023 season matches of the five major European national football leagues (England, Italy, Spain, Germany, and France). The original open-source dataset, provided by , contains many statistics for each match, including the average value of the pre-match odds for the following outcomes: Home team win (AvgH), Draw (AvgD), Away team win (AvgA), More than 2 goals scored (AvgO), Maximum two goals scored (AvgU). In addition, the information needed to identify a particular match (Div, Date, Time, HomeTeam, AwayTeam) and the information related to the final result (Full Time Result (FTR), Full Time Home Team Goals (FTHG) and Full Time Away Team Goals (FTAG)) were also included in the custom dataset *odds23.csv*.

Once the information from the 1826 matches had been collected, the MDS algorithm was applied to the five columns relating to the odds in the data pre-processing phase, thus obtaining the (x,y) coordinates for the scatter-plot. Figure 1 shows the initial visualization once loaded the project.

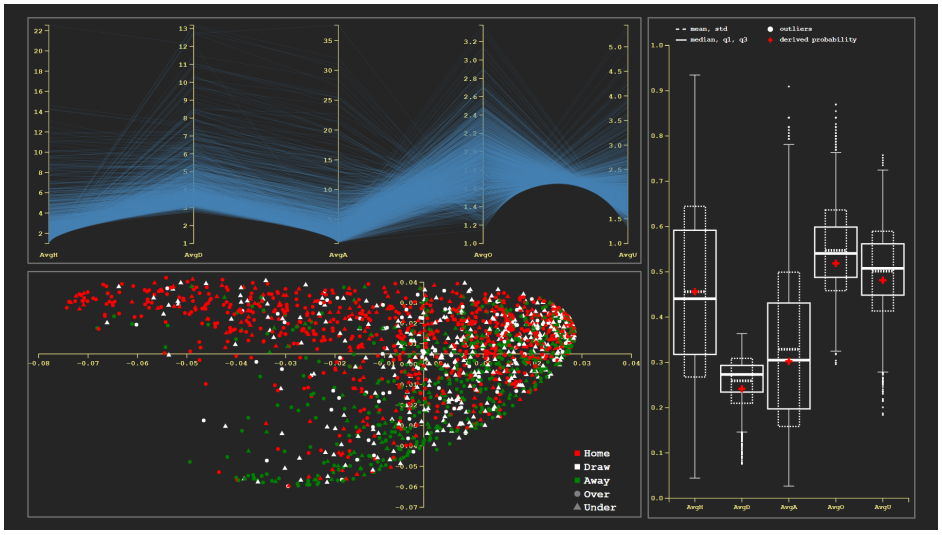

Figure 1: Dataset Overview - on the top left of the figure the parallel coordinates plot, on the bottom left the scatter-plot and on the right the comparative chart with b-boxes

4.1 Insights

The first information that can be observed by analysing the comparative chart and the b-boxes for the entire distribution is that these distributions well approximate Gaussian distributions. By Analysing the two-dimensional projection in the scatter plot, three clusters were identified that provided insights into the respective distributions.

The first cluster of data that stood out was located in the upper left corner of the scatter-plot, as shown in Figure 2. For these points, the bookmakers overwhelmingly favour the home team's win and the Over 2.5. The actual occurrence of these outcomes is in line with the probability derived from the distribution of these odds, as we expected. However, it is interesting to note that the underdog event, particularly the Under 2.5, appears in the distribution less frequently than what derived from the odds. This means that, according to our metrics, betting on Under 2.5 in these cases offers a negative expected return.

The second cluster analysed is the one on the far right of the scatter plot, which has the highest density of events, as shown in Figure 3. In this case, the bookmakers have no clear favourite, just a slight preference towards the home team, which is understandable given the advantage of playing in their own stadium. In terms of expected goals in the match, the under is slightly

favoured with respect to the Over 2.5 for these events. However, the position of the actual occurrence of the Over 2.5 outcome is unbalanced compared to the distribution, which may suggest that in this type of event, bookmakers tend to overestimate the probability of the teams scoring at least 3 goals.

The third cluster analysed is the one at the bottom, as shown in figure 4. The events that belong to the selected area are those in which bookmakers give the away team and the Over 2.5 outcome as favourites. Of all the clusters analysed, this is the one with the sharpest discrepancy between the actual occurrences of the outcomes given as favourites and the probability distribution derived from the odds offered by the bookmakers. According to our metrics, bookmakers tend to overestimate the probability of the away team winning and the amount of goals scored in events with these odds.

5 Conclusions and Future Work

This study presented a comprehensive exploration of betting odds through the development of an interactive web application, leveraging visual analytics tools and dimensionality reduction techniques. The coordinate utilization of Parallel Coordinates Plot, Scatter Plot with MDS, and Comparative Chart, facilitated a novel approach to analyzing sports betting data. Experimental results underscored the capability of the proposed system to identify clusters of betting odds that offer advantageous or disadvantageous opportunities relative to their derived probabilities.

Future Perspectives Looking ahead, this project opens several avenues for further research and development. The scalability of the application can be enhanced to accommodate larger datasets and a larger set of odds per match, enriching the analysis with a broader spectrum of betting opportunities. Additionally, incorporating machine learning algorithms for automatic cluster identification and prediction models could offer users insights into potential outcomes based on historical data. Another promising direction involves expanding the interactivity of the tool, enabling users to customize the analysis by uploading their datasets, applying different dimensionality reduction techniques or potential alternative methods for probability analysis.

In summary, the exploration of visual analytics in sports betting, especially within the context of football, reveals a landscape where significant opportunities for innovation and deeper insights exist. The current study

positions itself within this landscape, aiming to further the understanding of betting odds analysis through visual analytics and to pave the way for future advancements in this domain.

References

[1] Vishvak Bhatt, Udgam Aggarwal, and CNS Vinoth Kumar. Sports data visualization and betting. In *2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–6. IEEE, 2022.

[2] Rene Cutura, Christoph Kralj, and Michael Sedlmair. Supplemental material for druidjs—a javascript library for dimensionality reduction.

[3] Rene Cutura, Christoph Kralj, and Michael Sedlmair. Druid js—a javascript library for dimensionality reduction. In *2020 IEEE Visualization Conference (VIS)*, pages 111–115. IEEE, 2020.

[4] Richard C. Roberts, Robert S. Laramee, Gary A. Smith, Paul Brookes, and Tony D'Cruze. Smart brushing for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1575–1590, 2019.

[5] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):241–250, 2017.

[6] Nitin Singh. Sport analytics: a review. *learning*, 9:11, 2020.

[7] Erik Štrumbelj. A comment on the bias of probabilities derived from betting odds and their use in measuring outcome uncertainty. *Journal of sports economics*, 17(1):12–26, 2016.

Appendix

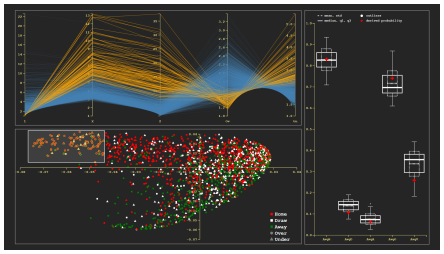

Figure 2: The points selected in the scatterplot are analyzed in the comparative chart and the relative odds are highlighted in the scatter-plot

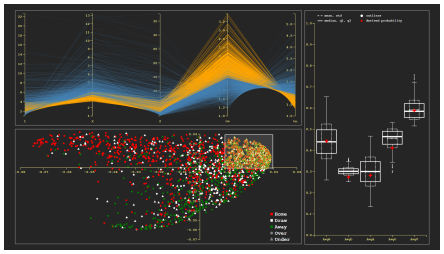

Figure 3: The points selected in the scatterplot are analyzed in the comparative chart and the relative odds are highlighted in the scatter-plot

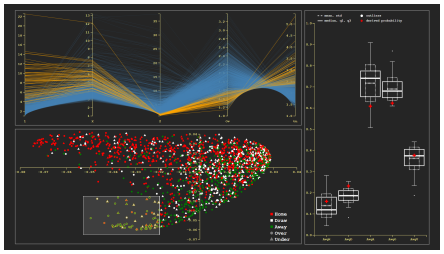

Figure 4: The points selected in the scatterplot are analyzed in the comparative chart and the relative odds are highlighted in the scatter-plot