NOVA IMS Information Management School

MDDM

Master's Degree Program in

Data-Driven Marketing

Predictive Modelling of Merchandising Sales in a Football Club

Project Work

Marta Francisco da Cunha Mendes Carneiro

presented as partial requirement for obtaining a Master's Degree in Data-Driven Marketing

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Predictive Modelling of Merchandising Sales in a Football Club

by

Marta Francisco da Cunha Mendes Carneiro

Project Work presented as partial requirement for obtaining the Master's degree in Data-Driven Marketing, with a specialization in Data Science.

Supervised by

Carina Albuquerque, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisboa, 10/01/2025

Marta Carneiro

DEDICATION

I would like to dedicate this project first and foremost to my family, who always supported me in my education and believed in me, thus supporting me throughout this great challenge.

I would like to thank all my friends who were there when I felt most unmotivated and who encouraged me to complete this project. As my friend Isis said, I was the only one who doubted myself.

I would also like to thank my work team who gave me the opportunity to develop this project and for the trust they placed in me to do so.

Thank you to my advisor, Carina, for making this project possible, no matter how impossible it may seem.

And finally, I would like to dedicate this project to my little star.

ABSTRACT

This project develops a weekly merchandising sales forecasting model for a professional football club with the goals of maximizing order quantity from suppliers, avoiding stockouts, and reducing overstock. Based on the CRISP-DM process, historical sales, performance of the club, and weather information were collected, cleaned, and analysed. The extracted variables were used to depict product launches and promotions, so that the last dataset could be weekly aggregated and employed to train and compare the various forecasting models: SARIMAX, XGBoost, LightGBM, and Random Forest. XGBoost performed better than the other models, exhibiting the following performance, RMSE of 84.221, MAE of 46010.88 and an adjusted R^2 of 0.846, being superior in detecting non-linear relationships and intricate patterns in the data. This study demonstrates how machine learning methodology can become a major value driver of operational efficiency, enabling inventory management and creation of more strategic marketing campaigns, in addition to maximizing fan experience through access to most desirable products.

KEYWORDS

Machine Learning; Forecasting Demand; Time series; ARIMA; XGBoost; Random Forest; LightGBM

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AA	Always Available
AIC	Akaike Information Criterion
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
B2C	Business-to-Consumer
CRISP DM	Cross-Industry Standard Process for Data Mining
DNN	Deep Neural Network
GBDT	Gradient Boosting Decision Tree
GOSS	Gradient-based One-Side Sampling
LightGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ME	Mean Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
NBA	National Basketball Association
PRCP	Precipitation
PRES	Sea-level Pressure
RBFN	Radial Basis Function Network
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with eXogenous regressors
SHAP	SHapley Additive exPlanations

SVM	Support Vector Machine
TAVG	Average Temperature
TF-IDF	Term Frequency-Inverse Document Frequency
TMAX	Maximum Temperature
TMIN	Minimum Temperature
TSUN	Daily Sunshine Duration
UEFA	Union of European Football Associations
VAT	Value Added Tax
WDIR	Wind Direction
WPGT	Peak Wind Gust
WSPD	Winds Speed
XGBoost	eXtreme Gradient Boosting

1. INTRODUCTION

Merchandising has become an integral aspect of football clubs' operations, transforming from a peripheral revenue stream to a core element of their commercial strategy. As global brands, football clubs generate millions of dollars annually through the sale of merchandise, including team kits, apparel, memorabilia, and lifestyle products (Gregório, 2021). This dynamic revenue stream not only supports the financial sustainability of clubs but also strengthens their relationship with fans, fostering a sense of belonging and loyalty (Stroebel et al., 2019). However, accurately predicting merchandising sales remains a challenge due to the complex interplay of factors influencing consumer behaviour in this context (Okeleke et al., 2024).

During a one-year internship starting in October 2024 with the brand activation team of a Portuguese football club, the challenge arose of predicting the quantity of merchandising products to be ordered from suppliers on a weekly basis.
The objective of this forecast is to place this order more precisely in order to avoid stock outages, improve the club's supply logistics and allow the creation of more direct promotions and marketing campaigns according to stocks of items that want to run out, for example.

Sales prediction is vital in supply chain management as it enables accurate alignment of production, inventory, and distribution with demand. By forecasting sales, businesses can prevent overstocking or stockouts, reducing costs and ensuring customer satisfaction (Nguyen et al., 2024). It helps manage risks, plan for seasonal trends, and optimize supplier relationships. Additionally, precise predictions enhance financial planning and resource allocation (Lipovetsky, 2022). Leveraging advanced analytics and predictive models, businesses can improve efficiency, responsiveness, and profitability, ensuring a competitive edge in today’s dynamic market environment (Bruckhaus, 2007).

The primary research goal of the project is to develop a forecasting model that can efficiently predict weekly merchandising products' sales for a professional football club. The prediction helps in more accurate ordering from suppliers to avoid stockouts and minimize overstock, thus enhancing the supply chain efficiency of the club.
Secondarily, the study expects to optimize inventory management by improved planning of logistics and actionable information for the generation of successful marketing campaigns.

To pursue this goal, the study tests and evaluate different forms of forecasting methods, most particularly machine learning algorithms, for the identification of the best and high-performing method in the context of sports merchandising, to be applied.

This research also demonstrates how predictive analytics not only improves operations performance but also the ability to improve fan experience by ensuring that there

are wanted products available and assisting in creating a more satisfying consumer experience.

The structure of this project is based on the CRISP-DM methodology, which guides the entire process through six key phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The study begins by identifying the business need - accurate weekly sales forecasting of merchandising for a football club - and proceeds with collecting and analysing data from multiple sources, including sales, weather, and match performance. The data is then cleaned, merged, and enriched with engineered features and different machine learning models, including XGBoost, SARIMAX, LightGBM, and Random Forest, are developed and tested. The project concludes with model evaluation based on predictive accuracy, in order to choose the best performing model, to be applied in the structure of the club's inventory management.

2. LITERATURE REVIEW

2.1 SALES PREDICTION

In the midst of uncertainty and intense competition, business leaders are finding out that old models alone cannot ensure competitive advantage. Sustainable growth cannot be achieved with traditional management strategies (Joel & Oguanobi, 2024). There is the necessity for the effective sales prediction and planning, which has a direct bearing on the financial aspects of businesses and how they carry out their operations in alignment with market demand. (Nguyen et al., 2024). Therefore, businesses are turning to data-driven approaches, leveraging vast amounts of data to make informed decisions (Eboigbe et al., 2023).

Predictive analytics, a subset of advanced analytics that uses historical data, statistical algorithms, and machine learning techniques, plays a vital role in forecasting future events (Bouallègue et al., 2024). Predictive analytics empowers businesses to make proactive decisions that foster growth and profitability by identifying emerging risks and opportunities, while also using historical patterns and trends to forecast future outcomes (Joel & Oguanobi, 2024).

2.1.1 BENEFITS OF INCLUDING SALES PREDICTION IN COMPANIES' MANAGEMENT

By including predictive analysis in its management, businesses have the opportunity to improve in the following areas:

Increased Profitability: Predictive analytics help businesses improve profitability through optimising pricing strategies, predicting demand, and repositioning marketing initiatives (Bruckhaus, 2007). For example, businesses can analyse historical pricing data and customer behaviour to develop pricing models that maximise revenue and profit margins (Assad et al., 2020). In addition, demand forecasting would assist in optimising production schedules, reducing inventory costs and minimising stockouts or excess inventory (Irfani et al., 2024).

Identify Opportunities: Predictive analytics can highlight growth opportunities by observing market tendencies (Joel & Oguanobi, 2024), segmenting consumer groups, and creating new markets and niches (Nur & Siregar, 2024). When doing this, businesses can capitalise on untapped markets and gain a first-mover advantage (Komolafe et al., 2024).

Risk Mitigation: Predictive modelling techniques and data-driven strategies help businesses anticipate future scenarios with much more precision (Sadana et al., 2023), minimising uncertainties related to expansion initiatives (Lipovetsky, 2022).

In essence, sales prediction and planning are not just about forecasting numbers but about making strategic decisions that drive success in an increasingly competitive and dynamic marketplace (Jackson et al., 2024).

2.2 ALGORITHM'S APPLICATIONS IN RETAIL SALES PREDICION

Machine learning (ML) algorithms are increasingly vital in the sports retail sector, serving as powerful tools for predicting sales and optimizing various business processes. The sources provided highlight different applications of ML, particularly focusing on demand forecasting for inventory and the impact of brand equity on sales performance.

A study developed by Selvakumar et al. (2024) presented an alternative solution for effective demand forecasting and inventory management in businesses, emphasizing how implementing a Seasonal Autoregressive Integrated Moving Average (SARIMAX) model with exogenous intercepts can streamline supply-demand forecasting and enhance inventory management procedures. The research aimed to elucidate univariate forecasting, the stability concept, ARIMA models, and seasonal ARIMA models with exogenous factors, particularly in the context of sales and stock price forecasting. The model was applied to a monthly sale of champagne dataset from 1964 to 1972. The study highlighted that the SARIMAX system increases accuracy in anticipating requirements by combining historical sales data with relevant external factors, providing valuable insights for strategic planning, production, distribution, and replenishment efforts. The overall system, built on machine learning algorithms, tackles the challenge of forecasting sales for seasonal items, aiming for high accuracy and reliability to empower businesses to optimize inventory control and maximize sales revenue.

Another study proposes an efficient and accurate SARIMAX model for sales forecasting in Business-to-Consumer (B2C) businesses, based on historical values and their seasonality, explicitly presenting it as a good choice for this purpose (Murugan et al., 2023). The primary goal of this sales forecasting system was to provide valuable insights for business decisions, resource allocation, and strategic planning to help businesses avoid overstocking or understocking situations, optimizing inventory levels to meet customer demand while minimizing carrying costs. The study utilized a historical sales dataset sourced from Kaggle or internal records, and considered SARIMAX an efficient model for sales forecasting because it specifically takes into account seasonality and trends present in sales data, it also demonstrated high accuracy and low error rates, with a Root Mean Squared Error (RMSE) value of 337.5 and a MAPE value of 6.97, which are quantitative indicators of higher forecasting accuracy and reliability.

A study focused on improving inventory management within a sportswear company's digital channel (Bastos, 2024). This project implemented tree-based regression models, specifically Random Forest, XGBoost, and LightGBM, to forecast demand for "Always Available" (AA) products. The aim was to optimize inventory levels by ensuring sufficient stock to meet customer demand without overstocking. The models were trained on a variety of data, including historical sales, inventory levels, website traffic, and promotional information. LightGBM was identified as the most robust method due to its superior performance with the

lowest RMSE. The project also showed that features such as stock levels, prices, website traffic, and previous sales figures were significant in predicting demand.

Another study, developed by Li and Wang (2024), explores the impact of brand equity on online sales within the sporting goods industry, using data from the Chinese e-commerce platform JD.com. This study used machine learning to predict sales performance, focusing on a brand score derived from sales data as a measure of brand equity. The study uses several machine learning algorithms, including Decision Trees, Random Forests, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Deep Neural Networks (DNN). A key finding was that brand equity, had a more significant impact on sales volume than price or customer review. Feature importance analysis further confirmed that brand score was the strongest predictor of sales. The research also looked at models using only brand equity features and those combining brand equity with textual data from product titles using Term Frequency- Inverse Document Frequency (TF-IDF). Interestingly, models using brand equity features alone generally performed better than those including TF-IDF features, demonstrating that brand equity itself is a robust indicator of sales performance. Specifically, Random Forest consistently showed strong results, with the best overall performance when using the combined features approach.

In order to enhance the competitiveness of retail enterprises, a sales forecasting approach was implemented grounded on the LightGBM framework—a new variant of the Gradient Boosting Decision Tree (GBDT)—to predict Wal-Mart sales (Deng et al., 2021). The objective was to provide accurate sales predictions to support strategic planning, including investing optimally, minimizing inventory expenses, maximizing profits and revenue, and minimizing risks. The model was trained on a Walmart dataset that had stores from California, Texas, and Wisconsin and along various dimensions such as time, product categories, department and item levels, store level details, and explanatory variables such as price, promotions, day of the week, and special events.

The forecasting model was trained on the first 1,413 days of data to predict sales over the following 28 days. Data preprocessing involved converting categorical variables to numerical and doing extensive feature engineering. LightGBM also demonstrated superior predictive power with a RMSE of 0.641, outperforming Logistic Regression (RMSE 0.803) and SVM models (RMSE 0.732).

Another research was designed to forecast food sales to enable the food industry to better plan and prevent wastage. The main goal was to identify the best machine learning methods to forecast sales in the food industry, specifically in a Swedish supermarket retailer, by performing data transformations and applying the algorithms in five various sets of sales data, each containing over 100 sales records. The research took into account other alternatives to forecasting sales, with a focus on Random Forest Regression and Gradient Boosting Regression specifically but it also mentioned other methods like Multilayer Perceptrons (MLP), Radial Basis Function Networks (RBFN), SVM, and various regression and time series models. The

models were compared on their accuracy score, mean absolute error (MAE), and max error (ME) and the outcomes clearly showed that Random Forest Regression was superior compared to other approaches, notably Gradient Boosting Regression (Naik et al., 2022)

In "Intelligent Sales Forecasting System Using Arima, Sarima, and XGBoost Models" by Atanda et al. (2024), the research aimed at developing an intelligent sales forecast system to predict future revenue by estimating product or service quantities, thereby supporting strategic decisions like pricing and inventory control. The authors identified that most companies are saddled with departmental silos and inadequate sharing of sales forecast data, resulting in misalignment in demand planning and with improper sales targets. As a response to all this, the system compared and merged three machine learning and artificial intelligence techniques: Autoregressive Integrated Moving Average (ARIMA) model, Seasonal AutoRegressive Integrated Moving Average (SARIMA) model, and Extreme Gradient Boosting (XGBoost) algorithm. An 18-column and 9800-row Kaggle dataset comprising product category, sub-category, customer name, country, and state was used for model training. The process involved extensive data preparation including gathering, cleaning (deletion of missing values and outliers), normalization, feature selection/engineering, and k-fold cross-validation-based data splitting, which is critical for time series data. Hyperparameters such as n_estimators, learning_rate, and max_depth in XGBoost were also tuned, typically by grid search, for improved performance and preventing overfitting. ARIMA and SARIMA model estimation employed maximum likelihood estimation, with XGBoost being trained iteratively by building decision trees to reduce error through fitting more trees on the loss function gradients. Accuracy of the models was checked against RMSE, Mean Squared Error (MSE), and MAE. The results clearly indicated that the XGBoost model significantly outpaced the ARIMA and SARIMA models and possessed significantly lower values of RMSE, MSE, and MAE. This demonstrates that XGBoost provides more accurate and efficient sales predictions and is more able to determine the underlying patterns and seasonal trends in sales.

2.3 MERCHANDISING IN THE FOOTBALL WORLD

Merchandise sales represent a vital component of a football club’s commercial revenue and, for many clubs, are essential to maintaining financial viability (Gregório, 2021). Beyond their economic significance, merchandise sales also serve a strategic function within the club's broader business model (Breitbarth & Harris, 2008). Importantly, these sales are not just financial transactions—they are also a means through which fans express their identity with the team, deepening their emotional connection and loyalty (Stroebel et al., 2019).

2.3.1 MERCHANDISE AS A REVENUE STREAM

Merchandise revenues have grown sharply in recent years, contributing to the commercialization of soccer and reducing the role of gate revenues in club revenues (Simmons, 2001).

According to UEFA's report The European Club Finance and Investment Landscape (UEFA, 2023), the financial performance of European football clubs shows a strong correlation between overall revenue and earnings from kit and merchandising. The top 20 clubs in this category largely overlap with the highest earners in total revenue, albeit with some variation in ranking. Notably, several well-supported clubs also feature prominently. Among the top performers, Manchester United leads with €130 million in kit and merchandising revenue, followed by Paris Saint-Germain with €97 million, Manchester City with €73 million, and Borussia Dortmund with €54 million. Additionally, clubs with large, loyal fan bases—such as AFC Ajax, Celtic FC, Leeds United FC, and Eintracht Frankfurt—also rank within the top 20 for kit and merchandising income.

Just as ticket sales, broadcasting revenues, and prize money, merchandising sales contribute to the club's overall income and growth (Van Haaren & Van Den Broeck, 2014), fans feel motivated to buy products from their club and actually feel happy with the purchase, as they know they are actively supporting the club to become a bigger company capable of making the best decisions for its performance, without financial capacity being an obstacle, or at least a less significant obstacle (Ahn et al., 2013).

2.3.2 MERCHANDISE AS FAN ENGAGEMENT

Wearing team merchandise allows fans to express and visibly display their affiliation with their favourite club, strengthening their identification with the team and fostering greater loyalty (Stroebel et al., 2019). Beyond personal expression, merchandise also serves a social function by signalling team allegiance, helping fans recognize one another and form communities built around shared beliefs and attitudes within the football world (Hedlund, 2014).

This simple use of merchandising can be seen as advertising made to the club by the fan, both in traditional media such as television, and on various social networks. Football clubs themselves take advantage of this act and use it for advertising, enriching the relationship with the fan, for example, the #OndaVerde movement used by Sporting Clube de Portugal, a hashtag that shares photos of sporting fans using the equipment of your idols around the world.

2.3.3 MERCHANDISE AND SPONSORSHIP

Merchandising products, namely pieces of equipment, are also seen as a space for partner advertising. The report from UEFA (The European Club Finance and Investment Landscape, n.d.) notes that sponsorship and commercial revenue increased strongly in 2023, a 30% increase on pre-pandemic levels from 2019.

For top clubs, sponsorship and commercial revenue accounts for between 40% and 50% of total revenues and is a main source of imbalance between clubs. The top 20 clubs' sponsorship and commercial revenue increased by 16% in 2023. The report provides the examples of the commercial revenue of some top clubs in 2023 in the table 1:

Table 1: Commercial revenue of some top clubs in 2023

Football Club	Revenue
Manchester City FC	€371m
Paris Saint-Germain	€330m
Real Madrid	€324m
FC Barcelona	€294m
Bayern Munich	€289m
Liverpool FC	€266m

Note. Adapted from UEFA (2023)

The report also states that domestic companies dominate main shirt sponsorship across Europe, but the Big 5 leagues attract global firms for international visibility. Sixty-five percent 65% of main shirt sponsorships across Europe have remained the same as the previous season.

2.3.4 EVOLUTION OF MERCHANDISING

The sale of merchandise has been on the rise in the last decades to the point at which income from such sales is now one of the biggest cashflows of football clubs-more than those from match days and broadcasts.

Historically, merchandising was concentrated on the sale of wearables, but has now diversified into selling toys, accessories, and many more.

Growth of the sporting industry as a commercial concern increased the requirement for effective merchandising strategies aimed at increased profit and brand value.

2.3.5 CONCLUSION

Merchandising is a multi-faceted aspect of a football club's business. It is not only a revenue source, but it is also a vital tool for strengthening fan loyalty, boosting team identification, enhancing brand visibility and contributing to a unique event experience. By strategically managing their merchandising activities, understanding fan behaviour, and embracing innovation, football clubs can enhance both their financial success and their brand value. Clubs should strive to make their merchandise more accessible, sustainable, and appealing to fans worldwide, whilst also ensuring that they are managing their supply chains effectively to maximize profitability

2.4 MACHINE LEARNING IN THE SPORTS WORLD

Machine learning nowadays finds its place in sports, transforming many aspects of the industry (Bunker & Susnjak, 2022), from predicting sponsorship costs to forecasting match attendance and player performance, these data-driven methods bring new insights and decision-making tools beyond traditional statistical analyses.

Jensen et al. (2014) explore the determinants of sponsorship costs within the highly competitive athletic apparel industry using market intelligence. Their research employs a hierarchical regression procedure to investigate the influence of factors like property-specific characteristics and on-field performance, finding these to be significant predictors of costs. Interestingly, they found that market-specific factors such as population size and income, were non-significant in predicting sponsorship costs in collegiate athletics. This study also identifies potential agency conflicts in resource allocation towards properties near sponsors' headquarters, as well as evidence of overspending by challenger brands such as Adidas and Under Armour, in their attempts to compete with industry leader Nike. This analysis provides an analytical framework that allows managers to better forecast and predict sponsorship costs.

Mueller (2020) investigates tree-based ensemble methods, specifically random forest regressions, for both pre- and within-season attendance forecasting in Major League Baseball (MLB). The study predicts individual game attendance for all 30 MLB teams, using data from 2013 to 2014 for model training to forecast the 2015 season. It employs 37 within-season predictor variables, including pre-season and short-run variables, with the aim of improving on past studies that mostly use linear regression models. The research also identifies important predictors, highlighting the value of using random forests for multiple forecasting horizons. The author notes that it is common practice to use attendance and ticket sales as proxies for sports demand and uses the terms interchangeably. The study uses a combination of statistical software including R, RStudio and the packages Random Forest, party, dplyr and others.

Claudino et al. (2019) offer a systematic review of the use of artificial intelligence (AI) for injury risk assessment and performance prediction in team sports. The review identifies 11 AI techniques or methods applied across 12 team sports, with the most frequently used being artificial neural networks, decision tree classifiers, support vector machines, and Markov processes. The study notes that AI applications were more prevalent in soccer, basketball, handball, and volleyball. This review emphasizes the potential for AI to enhance decision-making in sports, with a call for further evaluation research to establish predictive performance of different AI techniques. The research also indicates that AI techniques have shown better performance metrics than traditional statistical methods for predicting injury risk and athlete performance.

For the prediction of game outcomes in professional basketball, specifically the NBA, a real-time predictive model has been developed by integrating the machine learning XGBoost and SHAP algorithms (Ouyang et al., 2024) This model provides valuable decision support for various stakeholders, including coaches, athletes, club managers, and sports bettors. The XGBoost algorithm, has demonstrated optimal performance in predicting NBA game outcomes across different periods (first two quarters, first three quarters, and full game) concerning metrics like AUC, F1 Score, accuracy, and precision as it effectively manages model

complexity and prevents overfitting through regularization. To enhance the interpretability of the model, the Shapley Additive exPlanations (SHAP) algorithm quantified the contribution of individual features and indicated the following variables as the most influential for the model's prediction field goal percentage, defensive rebounds, and turnovers

In another study, developed by (Baboota et al., 2019) Machine learning in English Premier League soccer was applied in match outcome predictive analysis and modelling as a multi-class classification problem with three results: Home Win, Away Win, or Draw. Feature engineering plays a central role in this research, as it involves the creation and extraction of informative variables that enhance the model's ability to predict match outcomes. Among these features are team performance ratings—such as Attack, Midfield, Defence, and Overall—which are typically expressed as differentials between the home and away teams. Additional features include statistical indicators like Goal Difference and in-game performance metrics such as Corners, Shots on Target, and Goals, all of which are smoothed over recent matches to reflect current form. Temporal features are also incorporated, including Streak and Weighted Streak, which capture the momentum or direction of a team's recent performance, as well as Form, a comparative metric derived from a team's outcomes relative to their opponents. Together, these engineered features provide a comprehensive and dynamic view of match-related factors, significantly enriching the predictive capacity of the model. Gradient Boost and Random Forest models have had promising performances in this area, the Gradient Boost model did exceptionally well, especially for the drawing modelling, which is the least likely occurrence. Regardless, despite these advances, machine learning models for football predictions have not yet produced predictions that consistently exceed professional bookmakers' regularly, partly since there has not been copious detailed data presented like injury information or the presence of such important players.

2.5 FACTORS THAT IMPACT MERCHANDISING SALES

Merchandising sales at a sports club are affected by a variety of interconnected factors and Weather conditions and temperature play a significant role (Appelqvist et al., 2016). Sunshine can influence consumer behaviour and spending, though the effect can be complex as an increase in sunlight may boost spending at lower temperatures, but this effect could reverse at higher temperatures. Additionally, specific weather events, like unusually hot or cold spells, can impact consumer behaviour and spending in similar ways to holidays. Also, poor weather on match days can directly reduce attendance and merchandising sales. It's important to note that the quality and completeness of data is important for effective sales forecasting and planning. For example, data on weather conditions may sometimes be missing.

The timing of events is also crucial. Although one study did not find significant links between ticket sales and the day of the week or starting time, the day a match is played can still have an effect (Fotache et al., 2021). It is worth remembering that different studies may yield different results. Seasonality also plays a significant part, with some sports products being more affected by weather conditions and seasonal changes in demand (Appelqvist et al.,

2016). Seasonal businesses are more prone to disruptions caused by weather. Therefore, clubs need to consider the time of the year when planning inventory (Appelqvist et al., 2016).

In the paper written by Fotache et al. (2021), Team performance is also a notable factor. The home team's performance is generally more important than the away team in attracting fans to games and the team's recent form, measured by points gained in previous matches and their position in the league, can also influence sales.

Furthermore, research suggests that both the identity of the visiting team and the distance between stadiums can influence merchandise sales and related economic activity. High-profile visiting teams, such as globally recognized clubs, tend to attract larger audiences, which can boost merchandise sales and enhance the host team's brand image, as stadium visits are shown to significantly impact both team and city marketing efforts (Ramos et al., 2022).

Product-related factors are also important. In the paper written by Vashishtha et al. (2020), the time since a product was launched is a key consideration, with many fashion items having short lifecycles. It is also important to highlight that when a product is new, that is, it has just been introduced onto the market through a launch campaign, sales of this product tend to increase because it is new, also leading to an increase in other products that are sold in the same store or website. Furthermore, the attributes of a product like colour and sleeve length can influence sales, and these should be carefully considered when planning which items to sell.

Promotional activities and pricing strategies can influence demand as sales and selling prices have an inverse relationship, as the prices go up the demand decreases, so the increase in sales in any retail sector is closely linked to the activation of promotions on products and collections. Finally, external events, such as holidays and special occasions, significantly impact sports merchandise sales by creating spikes in consumer demand (Groebner,1990).

In conclusion, a sports club needs to be aware of weather, time of year, team performance, product characteristics and many other factors to successfully plan merchandise sales. By considering these interconnected influences, a club can improve inventory management and promotional activities and maximise its profits.

3. METHODOLOGY

The approach chosen for the development of this study was the CRISP-DM, method, as it has an iterative and flexible structure, ideal for solving business problems that involve data processing, analysis and evaluation (Chapman, 2000).

The CRISP-DM approach divides the whole process into six interconnected phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Gomez, 2020), allowing for efficient project management, maintaining focus on business objectives and enabling continuous feedback loops between phases, making it a cycle, as can be seen in the figure 1:

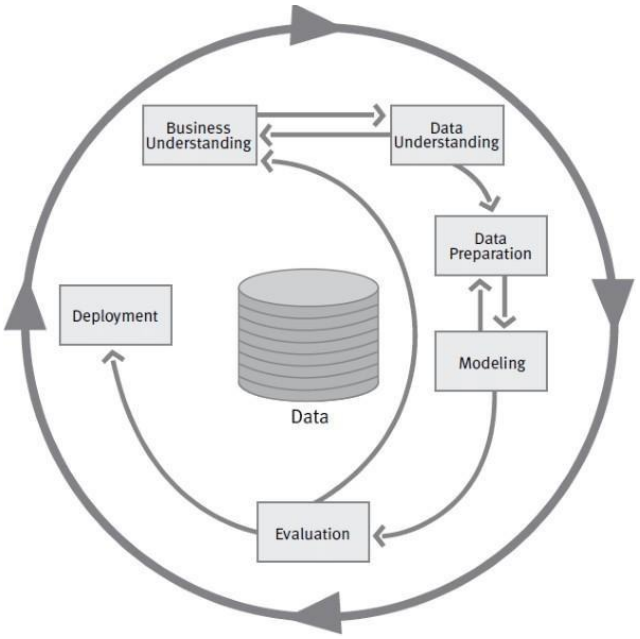

Figure 1 - Representation of the CRISP-DM model. Source: Wuttke (2023)

Furthermore, the technical success of the project significantly relied on the utilization of specific programming languages and development environments. The necessary historical sales data relevant to the problem was provided by the company, having been extracted from their internal data management and analysis platform, Power BI, and delivered in the format of excel files. Subsequently, the Python programming language was employed for data exploration, preparation of the received data, model development, evaluation, preparation of various graphs for data interpretation, transformations and analysis of model performance. All computational and analytical work was documented and executed within a Jupyter Notebook environment, which facilitated an interactive, and transparent approach by combining code, results, visualizations, and text, making it easy to document and share computational workflows in a readable and executable format.

The subsequent subsections describe the application of each phase of CRISP-DM within the specific context of this study.

3.1 BUSINESS UNDERSTANDING

The initial phase, Business Understanding, focused on clearly defining the business problem to be addressed: the forecasting of the total weekly sales value of football club's merchandise products. The primary objective was to develop a model capable of providing reliable estimates with a weekly granularity to support critical operational and strategic decisions.

This project opportunity arose from observed needs within the Merchandising, Marketing, and Logistics teams, particularly concerning stock forecasting and planning. Through discussions with the individuals responsible in these areas, it became clear that the existing method for forecasting these values was rather archaic and unsustainable in the long term, especially considering the current ordering processes in place.

A crucial aspect understood during this phase is the existence of two major categories of merchandise products, which are managed differently by the club due to having entirely distinct suppliers:

Technical Brand Products: These are products made in collaboration with the technical brand, one of the club's official sponsors, and provided by them. This category includes key items such as game kits, training wear, casual wear, and others. Ordering for these products is typically done only twice a year: in September and in January. These orders are intended to cover the sales for the entire season. As noted in the literature review, the football club merchandising market is highly dependent on the performance of the team itself. This performance is something that the merchandising and logistics teams cannot predict or control. Consequently, this limited ordering window and lack of flexibility throughout the year for technical brand products significantly disadvantages the business, particularly as the existing forecasting process relies simply on analysing sales from the previous season and adding or removing a margin based on the club's performance up to that point.

Own Brand Products: Created by the Merchandising team, these products embody the football club's brand DNA. They may occasionally be made in collaboration with other brands or simply feature the club's brand. These Own Brand products are gaining increasing importance and relevance, aligning with one of the club's major strategic missions encapsulated by the phrase "wear your club." The stock management for Own Brand products is more varied compared to the technical brand products, depending on the suppliers involved and potential partner brands for each item. The development of a more accurate and data-driven forecasting model, was thus identified as a critical need to improve efficiency, reduce stock-outs or overstocking, and enable better planning across the Merchandising, Marketing, and Logistics functions. It was essential to understand the needs of these various stakeholders and the potential positive impact that more accurate forecasts could have on optimizing the club's operations and maximizing merchandise revenue.

Identifying key factors known to influence sales, such as home/away matches, promotions, holidays, and seasonality, was also an integral part of understanding the business context and laying the groundwork for data analysis.

A major factor emphasizing the need for accurate forecasting is the club's track record over the years, because although it is positioned as one of the great clubs at national and European level, this club has gone through a difficult time on the field, which led to a decline in fan engagement, initially marked by decreasing supporter numbers, consequently resulting in a corresponding decrease in merchandise sales.

The club managed to overcome these bad seasons and has now returned to its original position associated with great achievements and victories. This turnaround led to the need for a new business model, which gave rise to this project as the club had already become used to the old turnover, and at the beginning it was difficult to adapt to these new sales numbers without a good analytical method associated with the company's business model.

This historical context highlights the sentimental nature of the sports merchandise business, as the bond between fans and their club is a major influential factor, directly impacting sales volumes and contributing to volatility in data, making accurate forecasting challenging yet critical (Habenstein et al., 2020)

3.2 DATA UNDERSTANDING

This phase involved gaining a thorough understanding of the available data, its characteristics, and its relevance to the forecasting problem. It encompassed the process of data collection and subsequent exploratory analysis and visualization to identify patterns, anomalies, and initial insights.

3.2.1 DATA COLLECTION

3.2.1.1 DATA COLLECTION - SALES DATA

The company provided the required historical sales data for the project, which was extracted from their internal data management and analysis tool, Power BI. and subsequently delivered in the Excel format. In total, 5 Excel files were delivered, each corresponding to the club's merchandising sales for each year, in accordance with the Portuguese sports calendar, i.e., the period for each file began on July 1st and ended on June 30th of the following year, except for the file relating to the 24/25 season, the year in which this project is being developed, before the season had finished, and therefore has a last date of March 17th, 2025.

3.2.1.2 DATA COLLECTION - CLUB'S PERFORMANCE DATA

Complementary to the club's sales dataset, data detailing the club's performance was collected through a manual process. This step was deemed essential based on the evidence from the literature review, which demonstrates a significant relationship between a football club's on-field performance and its commercial sales. The ZeroZero (n.d.) was utilized as the

source for extracting this performance information, which was then structured and stored in an Excel file.

3.2.1.3 DATA COLLECTION - METEOROLOGICAL DATA

To complement the primary transactional dataset and enable the exploration of potential external influences on the observed patterns, additional meteorological data was collected for the city where the club's physical merchandising stores are located because, as mentioned in the literature review, weather conditions can drastically affect a business's sales.

Daily weather information was obtained using the Python library meteostat, a tool that provides programmatic access to historical data from meteorological stations worldwide. In this process, it was necessary to define the location and dates (started on July 1st, 2020 and extended up to the date of June 1st, 2025) to obtain the corresponding meteorological data.

3.2.2 DATA ANALYSIS

3.2.2.1 DATA ANALYSIS - SALES DATA

As mentioned in the data collection phase, five Excel files relating to sales for each year were extracted. However, there were variables that did not exist in all data sets, due to the company's strategy and evolution over the years under analysis. Below, you can see Table 2 with the project variables, their description and their type.

Table 2: The sales dataset variables, their description and their type.

Variable	Type	Description
ARTIGO_ID	Categorical	Unique identifier for the transacted article (product/service)
SERIE	Identifier	Unique identifier for the point of sale
CHAVE_DOC_UNICO	Identifier	Unique key for the fiscal document associated with the transaction
DATA_VENDA	Date	Date on which the sale occurred
HORA_VENDA_CANAIS_FISICOS	Time	Time of the sale, specific to physical channel transactions

NOM_ENT	Categorical	Name of the entity associated with the transaction (e.g., customer, store)
QTD	Numerical	Quantity of the article(s) transacted
VALOR_LIQUIDO	Numerical	Net value of the transaction
VALOR_DESCONTO	Numerical	Value of the discount applied
VAL_SEM_IVA	Numerical	Value of the transaction excluding Value Added Tax
VALOR_CUSTO	Numerical	Cost value associated with the transacted article
VALOR_LUCRO	Numerical	Profit or loss value of the transaction
DESCR_SERIE	Categorical	Description of the point of sale
DOCUM_ID	Categorical	Short identifier for the fiscal document type
GENERICO	Identifier	Reference to aggregate several sizes of the same product
LOCAL_ROULOTE	Empty	Place of sale in Portugal of a trailer (mobile store)
NOME_ARTIGO	Categorical	Descriptive name of the transacted article
VALOR_CUSTONEW	Numerical	Cost value associated with the transacted article. Correction of VALOR_CUSTO

VALOR_LUCRONEW	Numerical	Profit or loss value of the transaction. Correction of VALOR_LUCRO
DOC_ORIGEM	Identifier	Identifier of the transaction's origin document
dia	Numerical	Day of the month for the transaction
Mês	Numerical	Month of the year for the transaction
local venda	Empty	Variable with no records
Origem	Categorical	Category representing the article's origin (related to the ID)
Família	Categorical	Category representing the article's family (related to the ID)
Tipologia	Categorical	Category representing the article's typology (related to the ID)
Negócio	Categorical	Whether the sale involves shipping costs or not
Sócio2	Categorical	Category related to a 'Partner'
Semana	Numerical	Week of the year for the transaction
Canal	Empty	Variable with no records

Variables like SERIE and CHAVE_DOC_UNICO have unique counts matching the total number of records, confirming their role as record-level or document-level identifiers. Others, such as ARTIGO_ID, NOME_ARTIGO, NOM_ENT, and DESCR_SERIE, show varying levels of granularity, with specific values occurring with high frequency, indicating dominant articles, entities, or document types within the dataset.

In contrast, variables like Negócio and Socio2 exhibit an extreme lack of variability. Negócio contains only the value '1 Portes' across all the records, while Socio2 only presentes the value '1 Sócio'.

Turning to the numerical variables, the statistics ('mean', 'min', 'max', '25%', '50%', '75%', 'std') offer insights into their central tendency, dispersion, and the presence of outliers or unusual values. A key finding is the presence of negative values in variables that typically represent positive quantities or monetary values, such as QTD, VALOR_LIQUIDO, VAL_DESCONTO, VAL_SEM_IVA, VALOR_CUSTO, VALOR_CUSTONEW, VALOR_LUCRO, and VALOR_LUCRONEW. These negative values strongly suggest the inclusion of records related to returns, credit notes or cost.

Turning to the temporal aspects of the data, DATA_VENDA is a fundamental and complete variable across all datasets, clearly defining the one-year period covered by each file. Derived temporal components like dia, Mês, and Semana (where present) show typical distributions consistent with calendrical patterns throughout the year.

Regarding data quality, several variables consistently suffer from significant missing data across multiple periods where they are present, indicating inherent incompleteness in their recording. HORA_VENDA_CANAIS_FISICOS is incomplete in all datasets, suggesting that transaction time is not consistently recorded for all sales channels or records. GENERICO and DOC_ORIGEM show high levels of missing entries in all periods where they appear, limiting their direct utility without imputation or exclusion. Tipologia also consistently has missing data in the later years.

Variables like local venda and Canal contain no records in all periods where they are included, indicating they are not populated in the source system for this data extract. LOCAL_ROULOTE is largely empty across all periods, containing data for only a limited number of records in the 2022/2023 dataset. This phenomenon occurred because during the summer of 2023, a club’s trailer travelled from north to south of the country in order to take the club's merchandise to anyone. Although the campaign did not continue in the following summers, the variable remained in the system and that is why it is empty from that year onwards.

Numerical variables, particularly the monetary values and QTD, consistently exhibit a skewed distribution where most values are concentrated at the lower end, with a tail extending towards higher values. They display a wide spread of values and a broad range between minimum and maximum recorded values, while the central bulk of the data (as represented by quartiles) is often concentrated in a much narrower band. This consistent pattern points towards the presence of atypical large-magnitude transactions that extend far beyond the common range across all periods. QTD consistently shows that most transactions are for a quantity of 1.

While many patterns are shared, each dataset presents a different number of records, reflecting different transactional volumes over the years, ranging from approximately 255,000 records in 20’21 to about 518,000 in 23’24.

In the datasets of 23'24 and 24'25, there are the variables VALOR_LUCRO and VALOR_LUCRONEW, together with VALOR_CUSTO and VALOR_CUSTONEW. This phenomenon occurred because, according to the teams, the variables with the word "NEW" came to replace the other alternative and were never removed from the system.

3.2.2.2 DATA ANALYSIS - CLUB’S PERFORMANCE DATA

As previously outlined during the data collection phase, a single Excel file was extracted, containing 254 records related to the club’s official football matches over several seasons. Each row represents a single game, described across eight variables that capture temporal, categorical, and performance-related information.

Table 3 summarizes the key variables, their type, and description:

Table 3 - Club’s performance data key variables, their type, and description

Variable	Description	Type
Data	Date of the match	Date
Hora	Kick-off time	Object (str)
Adversário	Opposing team	Categorical
Casa	Home (‘C’) or away (‘F’) indicator	Categorical
Resultado	Match result: ‘V’ (win), ‘E’ (draw), ‘D’ (loss)	Categorical
Pontos	Match points awarded (3/1/0)	Numerical
Pontos dos Últimos 8 Jogos	Sum of points earned in the previous 8 matches	Numerical
Competição	Name of the competition	categorical

To complement the statistical findings, a series of visualizations were developed in order to enhance an understanding of the club performance across time, across match conditions, as well as competitions.

The first graph shows the frequency of match results: wins (V), draws (E), and losses (D). The high number of wins proves the good performance of the club during the period in consideration.

The second diagram illustrates the cumulative points in the course of time. The diagram enables us to measure the performance of the team year after year and calculate the periods of stagnation or improved performance. The overall trend is always upwards, illustrating consistent achievement over the course of a few seasons. There are steep spikes in performance, particularly in late 2021 and for most of 2024, showing good match performances and high scores at these times. In contrast, we see flatter segments in early 2023 and mid-2025, indicating less intensified point accumulation, perhaps because of more losses or draws for those periods. The third chart is a boxplot of the average points scored per game across different competitions.

According to visual data, the team performs best in domestic leagues like Liga Portugal and Taça de Portugal, where points per game are nearly at the highest value of 3. In international league events like Liga dos Campeões or Champion's League, performance varies and is lower in general, and some games have zero points. The Super Taça also indicates widespread variation in performance, perhaps since it was played for just a few matches. The final graph traces the total points achieved during the last eight matches and clearly shows dips in performance over the period under consideration. For example, there are several highs to the peak of 24 points, indicating eight consecutive wins, noticeably in late 2021 and again in 2024. On the other hand, massive drops below 10 points are noticed in early 2023 and mid-2025, indicating times of low performance or challenging games; this is also confirmed with the second chart observations.

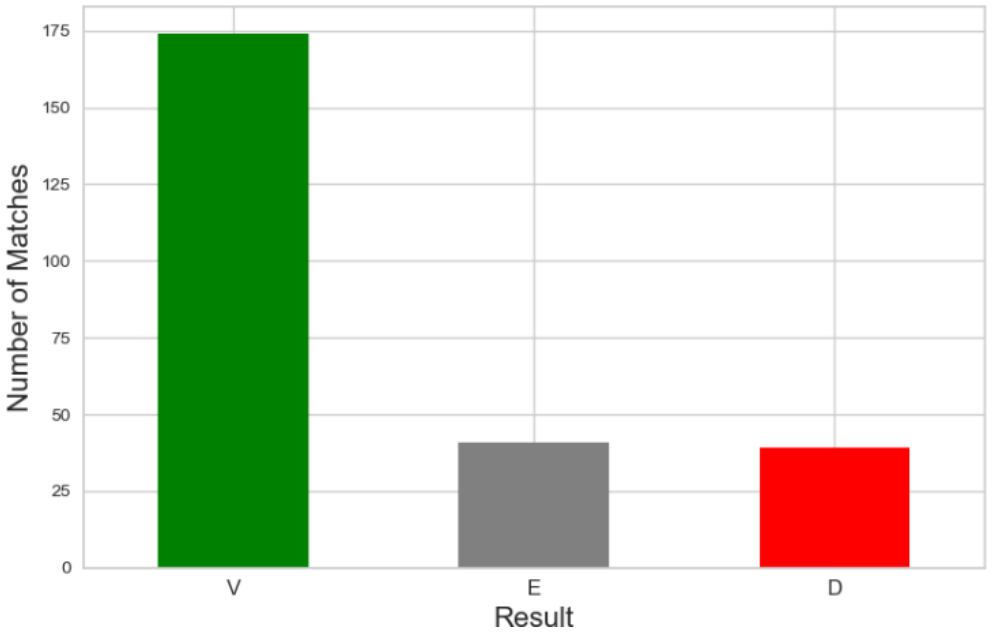

Figure 2 – Match Outcome Distribution

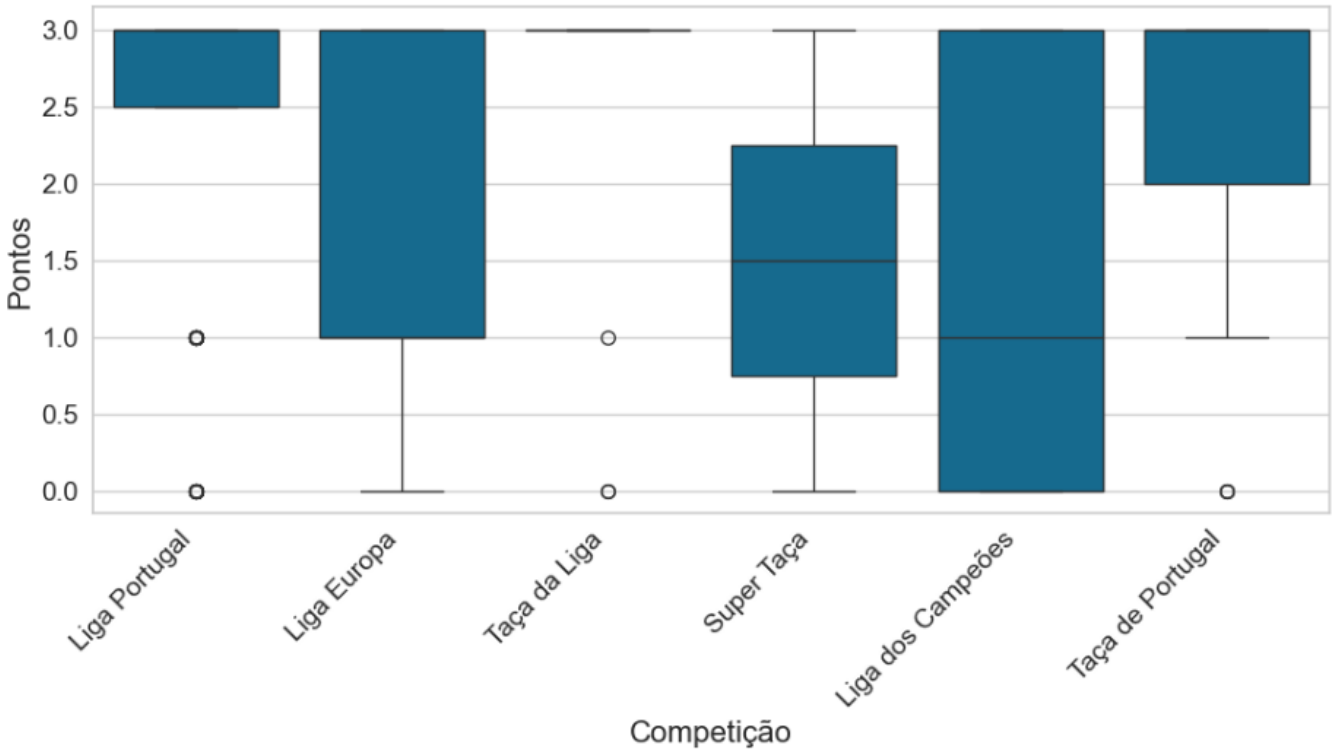

Figure 3- Points per Match by Competition

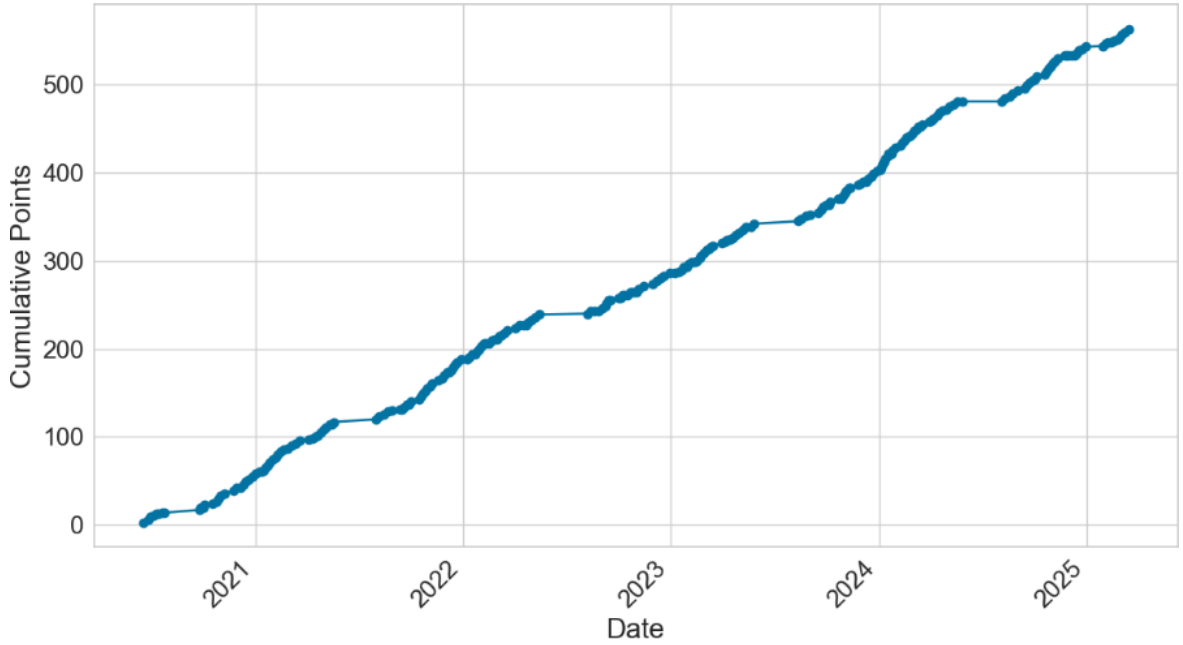

Figure 4 – Cumulative Points Over Time

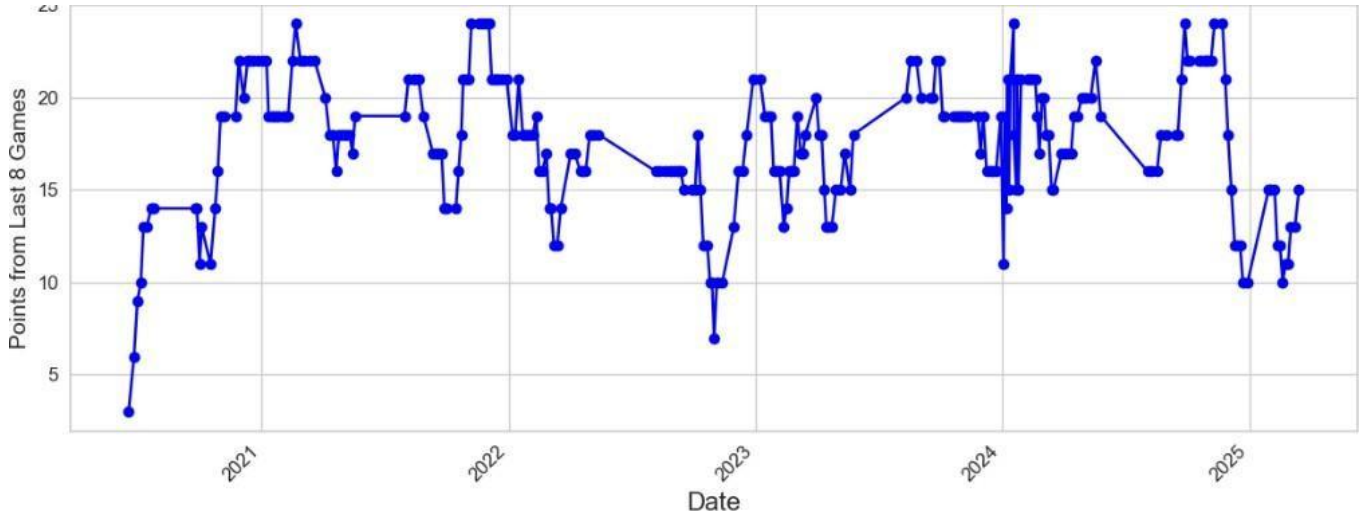

Figure 5 – Rolling 8-Match Points (Form Tracker)

3.2.2.3 DATA ANALYSIS - METEOROLOGICAL DATA

This dataset contains daily meteorological observations, with each record corresponding to the observations of a single calendar day, in the period from July 1, 2020, to March 17, 2025, comprising a total of 1721 entries and nine primary weather-related variables: average temperature (tavg), minimum temperature (tmin), maximum temperature (tmax), precipitation (prcp), wind speed (wspd), daily sunshine duration (tsun), snow, peak wind gust (wpgt), sea-level pressure (pres), and wind direction (wdir).

The variables tavg, tmin, and tmax follow a consistent and coherent pattern, tmin is always lower than or equal to tavg, which in turn is lower than or equal to tmax. The daily ranges suggest a temperate climate with an average daily temperature around 17°C. An outlier inspection reveals occasional colder days near freezing as well as warmer periods exceeding 35°C in summer.

The prcp variable stands out for its high frequency of zero or near-zero values, which is consistent showing a strongly right-skewed distribution. A small fraction of entries was missing, which were addressed using stochastic imputation based on the existing distribution which, preserves the realistic skew of rainfall occurrences while avoiding artificial uniformity.

Wdsp maintains a steady mean of around 15 km/h, with moderate variability. The distribution is approximately normal with mild skew, suggesting relatively consistent breezy conditions across the city throughout the year.

Both the tsun and snow variables lack any recorded data across the dataset, which indicates an absolute absence of data on these factors.

The wpgt variable shows a wide range of values from 11.1 km/h to 74.0 km/h max and a reasonably high standard deviation, which means high variation and the prevalence of extreme gust events.

The pres field is filled and shows a quite tight distribution, varying from 989.8 to 1036.9 hPa, pointing to atmospheric pressure stabilization across the period of observation.

Climate variability knowledge is required to identify seasonal trends that may affect customer behaviour, event planning, or operations planning (Bulgakova, et al, 2024) and for that reason some of the graphs were created which may reveal temperature trends during the period under investigation.

The highest chart (figure 6) summarizes the temperature and rain seasonal pattern in the region of the principal store of the club. Highest and average temperatures go on increasing from January to July and August at around 30°C (tmax) and 23°C (tavg), before falling towards December. Rainfall, however, is inversely patterned with summer values being lower (especially June to August) and highest in winter, especially November and December.

Finally, the heatmap (figure 7) illustrates average daily temperatures for each day of the month. The colour gradient serves to identify the hottest time of the year — late June to mid-August — when average daily temperatures often exceed 24°C. It is the coldest from December to February with average daily temperatures around 12°C. This chart adds more detail through the illustration of intra-month changes and smoothing season transitions.

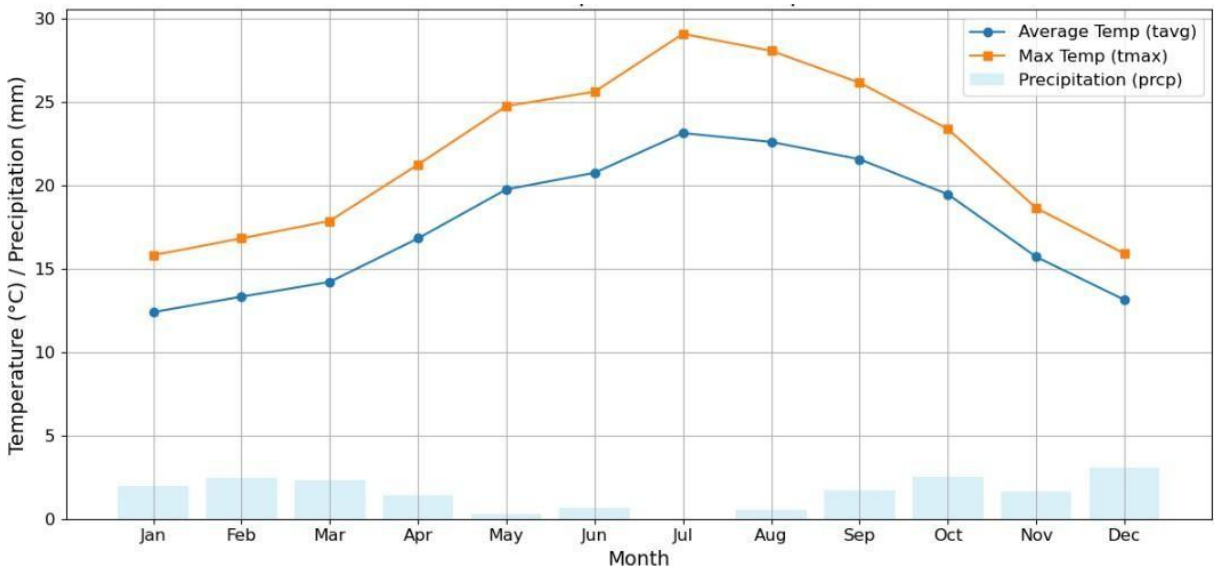

Figure 6 – Daily Average Temperature by Month

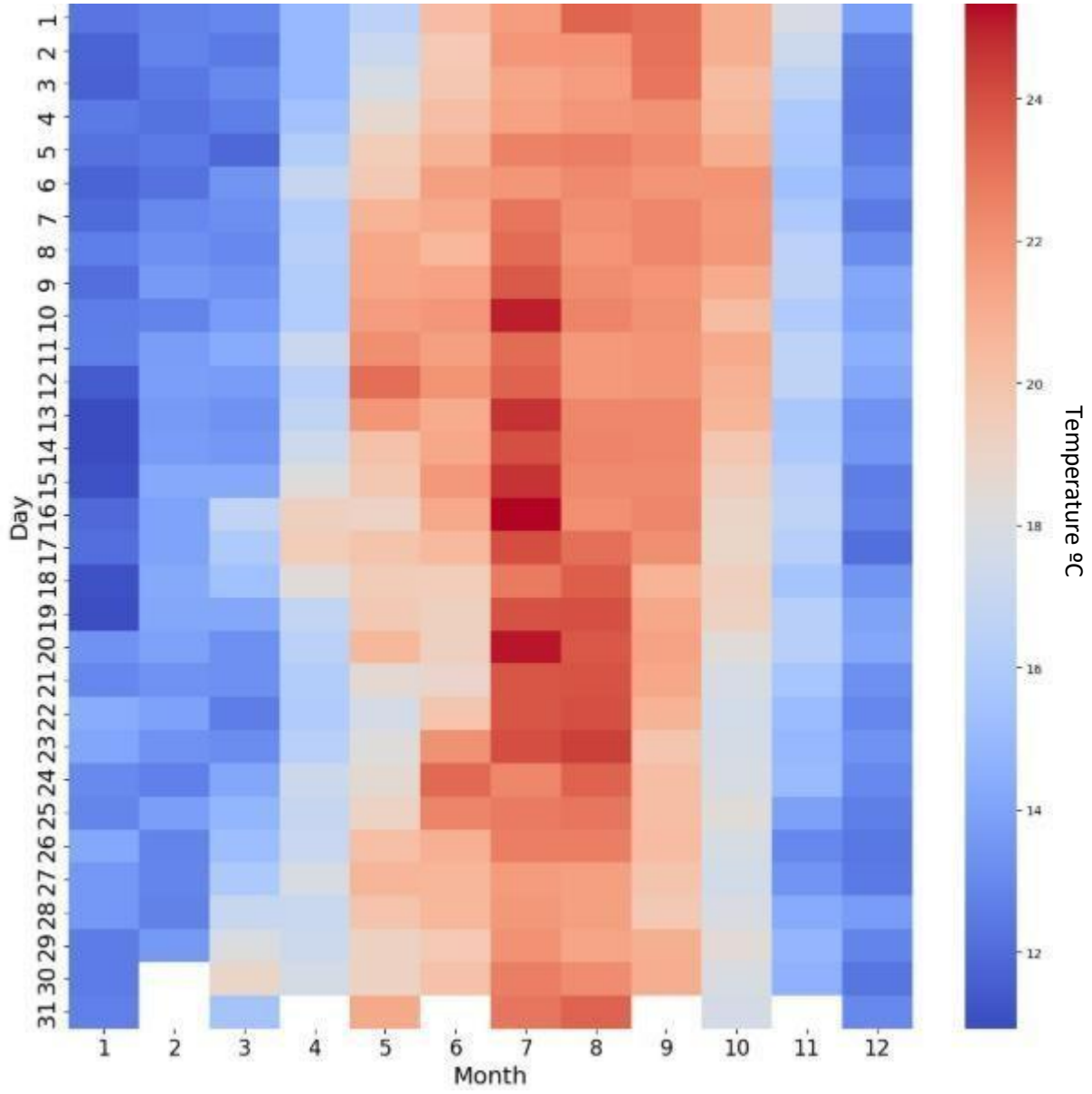

Figure 7 – Daily and Monthly Thermal Profile of the club's city

3.3 DATA PREPARATION

Following the Data Understanding phase, the Data Cleaning process was conducted as an intrinsic part of Data Preparation. This process was applied to all 5 sales data files and meant to solve identified data quality issues, more importantly regarding missing values, variability of variable presence across the datasets, redundancy of variables, and outliers' treatment, thereby preparing data for further analysis and model development.

3.3.1 DATA CLEANING

3.3.1.1 DATA CLEANING - SALES DATA

The initial assessment revealed several variables with significant data incompleteness. Variables identified as entirely empty across all datasets in which they appeared, specifically Canal and local venda, contained no records and thus offered no informational value and were removed. For the same reason, the LOCAL_ROULOTE variable was also removed, as it was largely empty across all periods, containing data for only a limited number of records in the 2022/2023 dataset, related to a specific campaign. Further variables exhibiting a considerable volume of missing data were also evaluated, such as DOC_ORIGEM, GENERICO, HORA_VENDA_CANAIS_FISICOS, SOCIO, and Negócio. These variables were either not consistently present across all datasets or did not possess an underlying logical structure or formula linking their values to other variables, which made reliable imputation or synthesis of missing values infeasible, so they were removed from the datasets in which they appeared. Similarly, in the 2022/2023 dataset specifically, the variables CATEGORIA and OPERADORCAIXA were removed due to analogous reasons of high missingness and limited presence across the full temporal scope of the project, only appeared in the mentioned dataset.

Conversely, some variables with missing entries were deemed valuable and amenable to imputation: the Tipologia variable, which presented missing values in certain datasets, was successfully imputed by identifying that the missing values corresponded to the first five digits of the value present in the ARTIGO_ID column for the respective transaction; similarly, the NOME_DO_ARTIGO variable also contained some missing values, which were filled by referencing other records within the same dataset where the ARTIGO_ID was identical, ensuring consistency in article naming.

Beyond missing data, an analysis of the relationships between numerical variables using Pearson correlation matrix (figure 8) revealed instances of variable redundancy. Specifically, it was observed that the values in the VALOR_ILIQUIDO and VAL_SEM_IVA columns corresponded exactly across all records where both were present. As these variables contained duplicate information, the VALOR_ILIQUIDO column was removed from the datasets in which it appeared, retaining VAL_SEM_IVA to represent the transaction value before VAT, thereby reducing redundancy without loss of information.

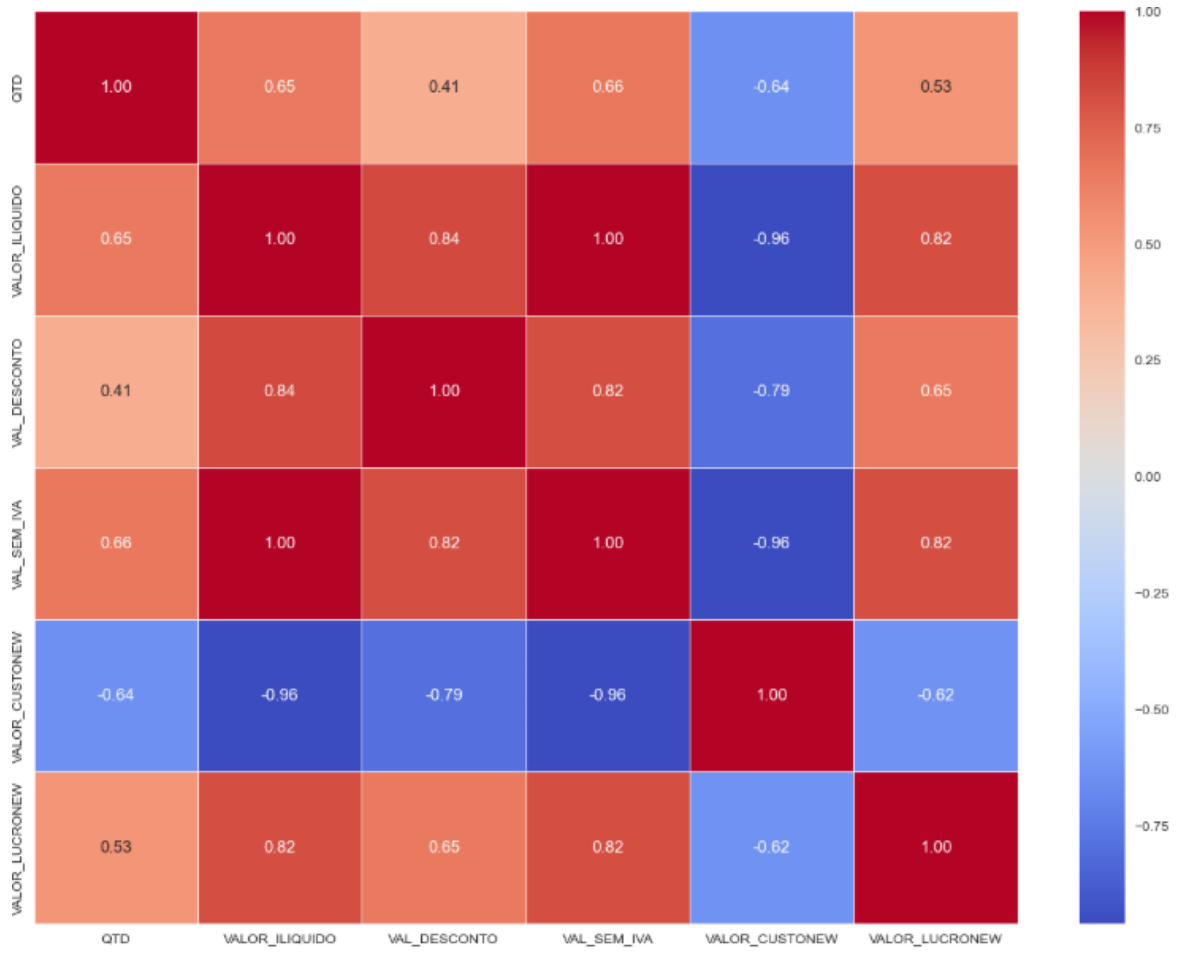

Figure 8 – Correlation matrix of the sales dataset variables

Regarding the treatment of outliers, the presence of extreme values and significant variability, particularly within the numerical variables such as quantity and monetary amounts were identified. However, recognizing that the underlying business experiences substantial oscillations and peaks, and that extreme values like large orders or significant returns are genuine events within the business context, a decision was made not to apply standard outlier removal or transformation techniques to most of these values.

The only exceptions to this approach were outliers specifically identified as being related to transactions with suppliers that were later returned (as discussed in the Data Understanding phase), as these were considered irrelevant to the primary objective of modelling customer-focused sales and were therefore excluded based on their source/type rather than solely their extreme value.

3.3.1.2 Data Cleaning - Meteorological Data

During the data cleaning process, the variables tsun and snow were removed as they were empty, not adding any new information to the dataset and wpgt, pres and wdir are more specific and are usually more important in studies focused on extreme weather events (Catto

& Dowdy, 2021). Since this study focuses on overall climate trends and seasonal patterns, they were also removed.
In addition, the prcp variable, which contained a smaller number of missing values, was cleaned through random imputation using observed non-null values, ensuring the original distribution of rainfall patterns was maintained.

3.3.2 FEATURE ENGINEERING

3.3.2.1 FEATURE ENGINEERING - SALES DATA

In this study, the Feature Engineering process primarily involved the creation of two new binary indicator variables.

The first variable created was LANÇAMENTO. This is a binary indicator variable designed to capture the anticipated increase in sales following the launch of key club kits. Based on historical launch dates provided by the company for each season, this variable was set to 1 for records occurring within the 15 days immediately following the launch date of the following kits: the Main jersey, Alternative jersey, 3rd Kit jersey, and 4th Kit jersey for each analysed season. Additionally, two specific kits launched in the 2024/2025 period were also included in this definition. For all other records outside these 15-day post-launch windows, the LANÇAMENTO variable was set to 0. This variable aims to represent periods of heightened customer interest and potential sales spikes.

The second variable engineered was CAMPANHA. This is also a binary indicator variable created to identify periods corresponding to major sales campaigns or seasons. Drawing upon historical data regarding specific promotional periods and sales seasons provided by the company, this variable was set to 1 for all records falling within these designated campaign periods for each year. For records outside these specific sales windows, the CAMPANHA variable was set to 0. This variable is intended to capture the effect of planned promotional activities specifically designed to stimulate sales volume and alter typical purchasing behaviour.

These two engineered features, LANÇAMENTO and CAMPANHA, translate specific, externally provided business initiatives (product launches and sales periods) into a format usable by the machine learning models. By explicitly marking these periods within the dataset, the models are enabled to potentially learn the distinct impact of these events on transactional behaviour, thereby enhancing the richness of the dataset for predictive modelling.

3.3.2.2 FEATURE ENGINEERING - CLUB'S PERFORMANCE DATA

In order to enrich the dataset, two key variables were engineered from the club's performance data: tipo_de_dia and importancia_competicao.

The first, tipo_de_dia, was created to place each calendar day in context in relation to the club's schedule of matches. For each match, a window of seven days was established—three

days prior to the event and three days subsequent—and each day was labelled based on its position in time and whether the match was at home or away. This yielded designations like Pré-Jogo (Casa), Dia de Jogo (Fora), and Pós-Jogo (Casa). Days that were not associated with any match window were labelled as Dia Normal. This aspect allows for temporal partitioning in subsequent analyses, with the ability to examine pre-match, during-match, and post-match patterns of behaviour.

The second characteristic, importancia_competicao, captures the perceived importance from each match by awarding points depending on the kind of competition, the home game factor with the presence of a classic opponent, and the location of the match. Higher scores are assigned to international competitions and rival matches played at home. This score was then aggregated at the daily level and merged into the main dataset, allowing each day to reflect the intensity or visibility of the various matches.

3.3.3 DATA MERGE AND AGGREGATION

Following the Data Cleaning step, Feature Engineering was performed, being intertwined with necessary data integration and aggregation. Feature engineering is important in enhancing the predictability of machine learning models through the transformation of available variables into new ones expressing connected information or relationships and bringing together data from disparate sources in standardized forms suitable for various analysis objectives.

The first was to merge the five cleaned annual transactional datasets (covering the period 2020/2021 through to the half-year 2024/2025) together into a single complete dataset. This involved bringing together all rows from the cleaned dataset files into a single composite DataFrame containing all transactions over the entire period of study. This is the entire universe of transactions to be analysed.

For the objective of predicting the overall weekly sales volume across all articles, the comprehensive transactional dataset (including engineered features, game, and meteorological information) was aggregated to a weekly level. This involved summing numerical variables such as VAL_SEM_IVA, QTD, VAL_DESCONTO, VALOR_CUSTO, and VALOR_LUCRO for all transactions occurring within a given week. Binary variables like LANÇAMENTO and CAMPANHA were aggregated by calculating their mean value for the week, indicating the proportion of transactions under that condition. Categorical variables like Mês were represented by the mode value within the week. Game-specific information (number of unique games, total points, and details of the most important game) and weekly meteorological data were joined to this aggregated dataset based on the corresponding week. This process resulted in a single dataset where each row represents a week, containing total aggregated sales metrics and shared weekly exogenous features.

3.3.4 DATA TRANSFORMATION

Additionally, rolling averages were also calculated for means in windows of 3, 4, and 8 weeks (media_3s, media_4s, and media_8s) based on lagged values and provided smoothed estimates of recent patterns in sales. For even more accurately reflecting short-term fluctuation, the first-order difference (delta_1) and week-over-week percentage change (pct_change_1) were computed.

3.3.5 ENCODING OF CATEGORICAL VARIABLES

To prepare the dataset for use with the model, the categorical variables were one-hot encoded. Namely, the columns 'Competição', 'Casa', 'Resultado', 'Tipologia', 'NOME_DO_ARTIGO', 'DOCUM_ID', and 'Origem' were identified as being categorical and accordingly encoded. One-hot encoding was applied to these variables to translate every category into a binary column in such a way that the model can read categorical data in numerical form. The drop_first=True argument was used to eliminate multicollinearity by removing the first category of each variable, thus eliminating duplicate data in the final feature set.

In addition to one-hot encoding, label encoding was used to transform the 'Adversário' variable. This variable, which represents the opposing team, was considered categorical but with a potentially large number of unique values. For dimensionality efficacy, scikit-learn's LabelEncoder was utilized, whereby a single integer was assigned to each distinct opponent. This maintains the categorical domain of the variable but converts it to the numeric format that can be employed for model input.

To make it uniform and compatible with the machine learning algorithms, all features were then explicitly converted to float type. This conversion was necessary because certain models, such as XGBoost, perform numerical computations that require uniform data types. In addition, converting categorical indicators (originally uint8) to float avoids data type mismatches during model training and contributes to overall stability and efficiency.

Then, the final feature matrix was constructed by concatenating the numerical features—excluding the original target variable (VAL_SEM_IVA) and the previously encoded categorical columns—with the one-hot encoded dummy variables. This resulted in a comprehensive dataset that combines both engineered numerical features and encoded categorical variables.

Finally, the target variable (y) was defined as the original sales value without VAT (VAL_SEM_IVA), which the predictive model aims to forecast and by the end of all these transformations, the final variables are visible in annex 1.

3.4 DATA VISUALIZATION

The figure 9 illustrates the monthly revenue (excluding VAT) across the calendar year, showing peak sales in August and July. This spike is expected, as it coincides with the start of the

football season and the launch of new kits, periods typically associated with increased merchandising demand.

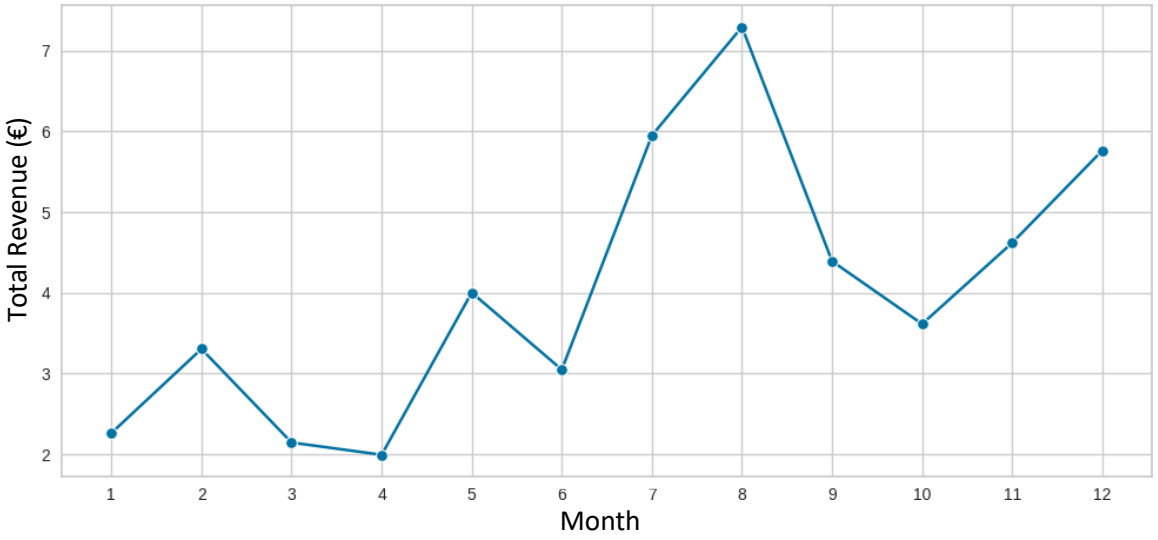

Figure 9 – Monthly Revenue (VAL_SEM_IVA)

The bar chart displays the top 10 best-selling articles based on quantity sold. "ESTAMPAGEM DE LETRAS" leads as the most sold item, followed by "Letras Principal" and "PORTES DE ENVIO." Notably, several items related to personalization and packaging—such as different types of paper bags and name/number printing—also appear among the top sellers. This trend is expected, as personalization items like name and number printing are commonly added across all jersey sales, making them a consistent and transversal component of merchandising purchases.

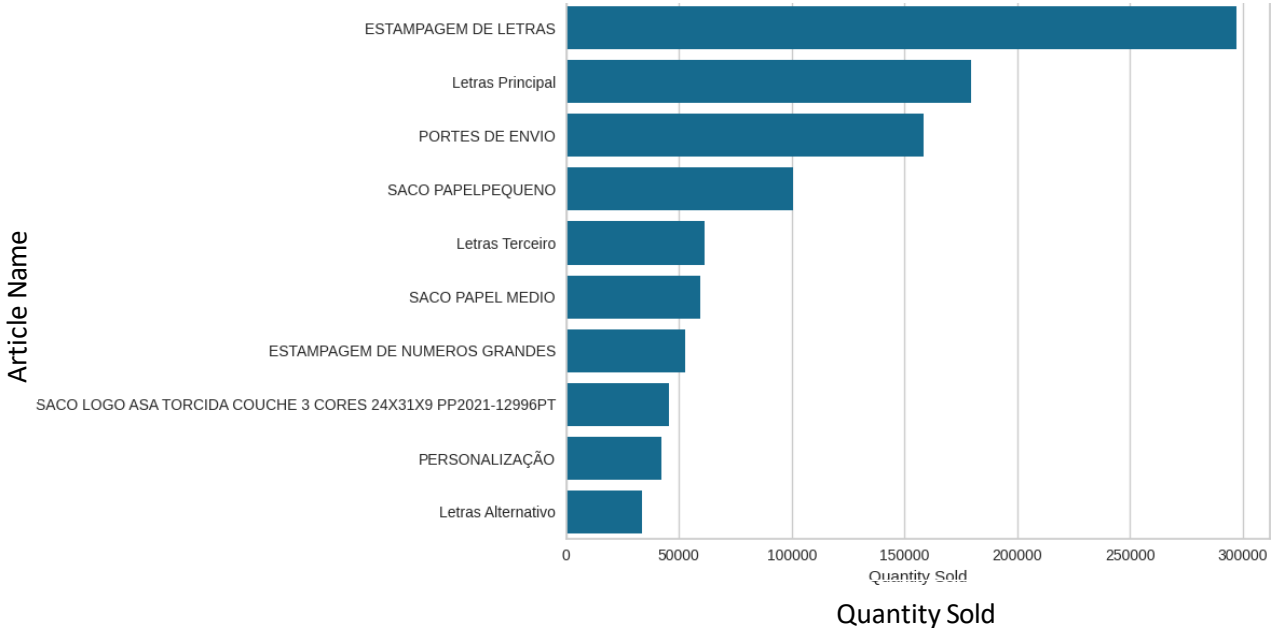

Figure 10 – Top ten Best-Selling Articles

3.5 DATA SPLIT

To train and test the XGBoost model, the dataset was initially split into training and testing datasets using the train_test_split function, with 80% of the data allocated for training and 20% for testing.
The shuffle=False parameter was passed to ensure that the chronological order of the observations is preserved, an essential step when working with time series data, to prevent data leakage from the future to the past.

3.6 MODELLING

In order to account for the temporal dynamics and external influences on the data, four modelling approaches were adopted: SARIMAX, Random Forest, LightGBM, and XGBoost. They were selected because of their complementary advantages as well as their ability to meet the challenges that are usually embedded in real-world forecasting problems such as seasonality, trend shifts, missing values, and complex, non-linear patterns (Benitez et al., 2023; Sheridan et al., 2016; Nafouanti et al., 2023).

SARIMAX extends the standard ARIMA model with the provision to add exogenous regressors. It is most effective in modelling seasonals as well as non-seasonals in univariate time series, specifically when there are domain knowledge or external variables which are assumed to affect the target series (Tseng et al., 2002). Although, its computation is costly in the scenario of most distinct time series as it requires training a model over every series.

In addition, XGBoost was also used as a machine learning-based solution. It is a robust tree-based ensemble method that is especially good at picking up non-linear relations and interactions among features. Unlike SARIMAX, XGBoost does not inherently need temporal ordering and therefore is well suited for using engineered features such as lag values, rolling statistics, and calendar variables. Further analyses, it is also very scalable and robust to outliers and irregular data distribution (Chen et al., 2020).

Two additional tree-based models were tested, LightGBM and Random Forest. LightGBM is a high-speed and high-efficiency gradient boosting framework most ideally suited for big data with high-dimensional features. It uses a leaf-wise tree construction policy and advanced techniques like Gradient-based One-Side Sampling (GOSS), which enables fast training and strong predictive capability (Xu et al., 2021).

Random Forest, on the other hand, is a bagging-based ensemble that builds multiple decision trees from bootstrapped data and random feature selection and gives high precision and highly resistant to overfitting (Karabadji et al., 2023). Both LightGBM and Random Forest support feature importance analysis and are applicable for structured data with complex, non-linear relationships (Yan et al., 2023)

3.6.1 XGBOOST

XGBoost (Extreme Gradient Boosting) is a tree-based model that is recognised for its impressive performance in machine learning and data mining tasks (Chen, 2023). Following the implementation of an internal method that combines the results from multiple individual trees, accurate predictions are achieved (Noorunnahar et al., 2023). This ensemble technique utilizes various methods to enhance the speed of decision trees and emphasizes lowering the computational complexity for identifying the optimal split, which is the most time-intensive aspect of decision tree building algorithms (Bentéjac et al., 2021). Furthermore, XGBoost can be used for feature selection, enhancing model interpretability and identifying key drivers in data by employing metrics such as 'weight' to rank feature importance, which aids in understanding the influence of different features on the model's predictions. (Davagdorj et al., 2020). The sources show that XGBoost is not just accurate, but also robust, and generalises well across different scenarios: it has been effectively used in various forecasting tasks, including heart disease prediction, flash flood risk assessment, and urban land use classification, demonstrating its adaptability and robustness across different domains (Georganos et al., 2018; Ma et al., 2021; Budholiya et al., 2020).

3.6.1.1 MODELLING – XGBOOST

The XGBoost estimator was created using the XGBRegressor class of the xgboost library. The model was initialized with 100 estimators, 0.1 learning rate, a maximum tree depth of 5 and a fixed random_state of 42 was specified to enable reproducibility of results. The model was trained on the training set predicted on the test set.

3.6.1.2 FEATURE SELECTION – XGBOOST

To gain insight into the contribution of each input variable to the XGBoost model's predictions, feature importance scores were extracted using the feature_importances_ attribute of the trained model and the variables QTD, "VALOR_LUCRO", and "VALOR_CUSTO" were highly correlated with the target variable and could cause data leakage, so they were removed, which ended up happening also before the application of the remaining models.

3.6.1.3 HYPERTUNING PARAMETERES – XGBOOST

To further improve the model's performance, the XGBoost regressor was re-applied using hyperparameter tuning through RandomizedSearchCV from the sklearn.model_selection module. A defined search space of key hyperparameters—including the number of estimators, learning rate, tree depth, and regularization parameters—was explored. The search was configured to perform 50 randomized combinations, using 5-fold cross-validation resulting in a total of 250 model fits, identifying the best-performing set of hyperparameters as follows: n_estimators: 500; learning_rate 0.1; max_depth: 3; subsample: 0.8; colsample_bytree: 1.0; gamma: 5; reg_alpha: 0.01; reg_lambda: 1.

To further refine the XGBoost model, a second round of hyperparameter optimization was conducted using Optuna, an efficient framework for automated hyperparameter tuning. A custom objective function was defined using 5-fold cross-validation with the negative RMSE as the scoring metric. A total of 50 trials were run during the optimization process and the best-performing set of hyperparameters found by Optuna was then selected to build the final version of the model.

3.6.2 SARIMAX

SARIMAX is a strong time series forecasting statistical model distinguished by its ability to model both the seasonal and the non-seasonal patterns and to incorporate external variables influencing the target series (Nontapa et al., 2020). By extending the SARIMA model to incorporate the addition of external variables, SARIMAX enhances the precision of predictions by compensating for known drivers external to the target variable itself (Elshewey et al., 2022).

The model works by first removing trends to stabilize the data, then using previous errors and previous values to achieve short-term patterns and finally combining seasonal components to capture repeating cycles. Its capacity for incorporating exogenous variables allows practitioners to add domain knowledge and external factors such as holidays, special promotions, or economic factors, which can have an important impact on the quality of the forecast. This is supported by studies in which the use of such variables yields improved predictions, and SARIMAX is superior to that of neural networks and traditional models (Ampountolas, 2021).

3.6.2.1 FEATURE SELECTION -SARIMAX´

To support the forecasting of weekly sales using the SARIMAX model, a preliminary feature selection process was carried out using Lasso regression with cross-validation. The objective was to reduce the dimensionality of the predictor space and retain only the most relevant explanatory variables. This was achieved through a modelling pipeline that combined data standardization using StandardScaler with Lasso regression (LassoCV), which automatically selected the optimal regularization parameter via 5-fold cross-validation. The model was trained on the complete set of predictors, and variables with non-zero coefficients were considered significant contributors to the prediction of weekly sales. A key strength of Lasso lies in its ability to perform automatic variable selection by shrinking the coefficients of less informative features to exactly zero. After fitting the model, only the predictors associated with non-zero coefficients were retained, representing those that most effectively explained the variability in merchandise sales.

The Lasso model selected the following 15 variables as the most important: NOME_DO_ARTIGO_PORTES DE ENVIO, Mês, Nº de jogos da semana, Tipologia_SHIPP, media_3s, VAL_DESCONTO, lag_4, Tipologia_90202, Tipologia_10123, NOME_DO_ARTIGO_PORTES DE ENVIO - NORMAL, Origem_2, LANÇAMENTO, VENTO, lag_2,

and Competição_Liga Portugal. All these features were subsequently used as exogenous variables in the SARIMAX model to improve its explanatory and predictive capacity.

3.6.2.2 MODELING – SARIMAX

Given the inclusion of multiple exogenous variables with varying scales, standardization was applied to the exogenous data using z-score normalization via StandardScaler.

To identify appropriate model parameters, the auto_arima function from the pmdarima package was employed. This automated the selection of the non-seasonal and seasonal components of the SARIMA model by optimizing information criteria while accounting for weekly seasonality (m = 52). The parameters returned by auto_arima were then used to specify and fit a SARIMAX model using the statsmodels library, incorporating the scaled exogenous regressors.

3.6.2.3 HYPERPARAMETER TUNING - SARIMAX

As the second modelling strategy, manual grid search was conducted to find out the optimal SARIMAX specification to be used in forecasting weekly sales (VAL_SEM_IVA). Unlike the initial strategy that utilized auto_arima for automatic selection of parameters, this different strategy utilized systematic sets of non-seasonal and seasonal ARIMA parameters based on Akaike Information Criterion (AIC).

To find the best-performing model, every combination of (p, d, q) parameters between 0 and 1 was attempted for both the non-seasonal and seasonal components, and a seasonal frequency of 52 weeks was employed to include yearly seasonality. All potential SARIMAX model contenders were then fit on the training set with normalized exogenous variables, and relative performance was compared based on the AIC. The model specification with the minimum AIC that was achieved by pooling orders was selected as the best model specification: SARIMAX(2, 0, 2)(0, 1, 0)[52] with intercept, i.e., the optimal model has two autoregressive and two moving average in the non-seasonal component, with the season difference being of order 1, without any seasonal moving average or autoregressive components.

3.6.3 LIGHTGBM

LightGBM (Light Gradient Boosting Machine) is a decision tree-based gradient boosting framework that has been shown to be extremely efficient and effective in the fields of classification and regression tasks (Seto et al., 2022). This model constructs trees leaf-wise, as opposed to the more conventional level-wise, allowing it to construct deeper trees, which leads to less loss and improved performance, especially in situations with complex, non-linear relationships.

LightGBM is also resistant to outliers and does not require extensive data preprocessing, like scaling or normalization, making it a robust and convenient solution for the majority of real-world use.

3.6.3.1 MODELLING – LIGHTGBM

To predict weekly sales for the merchandise store, the algorithm was used, specifically through the LGBMRegressor implementation. The model was trained using all available features, without performing prior feature selection.

3.6.3.2 FEATURE SELECTION – LIGHTGBM

Feature importances were subsequently obtained after training through the feature_importances_ attribute of the model. The importances reflect the frequency and success with which the usage of each feature was utilized to split the data over the trees, providing an indication of their contribution towards the predictive power of the model. The importances were converted into a Pandas Series, sorted in descending order, and the 15 most important features were printed.

According to the feature importances analysis, a subset of the most significant variables was selected manually to be used in an optimized version of the model. The selected features were: lag_1, delta_1, pct_change_1, VAL_DESCONTO, CHUVA, media_4s, VENTO, lag_2, LANÇAMENTO, lag_4, TEMP MIN, TEMP MAX, media_3s, media_8s, and Adversário. Then, training and test datasets were restricted to this handpicked subset of features, which was hoped to contain model complexity, increase interpretability, and potentially improve prediction performance.

3.6.3.3 HYPERPARAMETER TUNING – LIGHTGBM´

After identifying the most relevant features, a second model of the LGBMRegressor was trained using only the filtered variable set. To improve model performance and generalizability, a hyperparameter tuning process was conducted using the grid search method. The GridSearchCV scikit-learn function was employed with the already trained LightGBM model, together with a time series split in order to preserve the order of the data and avoid look-ahead bias. The grid search was optimizing the following hyperparameter combinations: learning_rate; max_depth; num_leaves; min_child_samples; subsample; colsample_bytree.

To improve predictive power, a grid search with time series cross-validation was carried out with 4 splits. The measure used during this exercise was MAE, negative scoring reversed for interpretability. This ensured that the optimal combination of hyperparameters for the LightGBM model could be set. The best-performing model was: colsample_bytree: 0.7; learning_rate: 0.05; max_depth: 3; min_child_samples: 10; num_leaves: 15; subsample: 0.7.

Using this tuned array of parameters, a new model of LightGBM was then developed. It achieved a MAE average of 56,023.33 across the validation folds, better than the baseline model that was trained on default parameters.

For additional optimization of the LightGBM model, Bayesian optimization using the Optuna framework was performed. Unlike grid search that thoroughly checks pre-specified sets, Optuna applies a more efficient approach to search wherein it learns to adjust its exploration of the space of hyperparameters based on previous results. The same cross-validation strategy with 4 splits was used in order to ensure temporal consistency when testing the models. The objective function minimized the MAE mean across the validation folds. The following hyperparameters were tuned: learning_rate, max_depth, num_leaves, min_child_samples, subsample, and colsample_bytree, while the number of estimators (n_estimators) was fixed at 1000. Early stopping in training was enforced in order to prevent overfitting.In order to guide the process of optimization, the best configuration found in the past grid search was used as a seed trial. After 50 trials, Optuna had discovered another group of hyperparameters reducing the validation error again: 'learning_rate': 0.0449205232637701, 'max_depth': 7, 'num_leaves': 62, 'min_child_samples': 12, 'subsample': 0.8747186692894985, 'colsample_bytree': 0.7056299713632427. The best configuration produced a mean cross-validated MAE of 50851.42, demonstrating a progressive improvement over grid search outputs.

After the hyperparameter tuning using Optuna, the best configuration was utilized to train the final LightGBM model. The final model was trained on the entire training data set and validated against the test set. Early stopping with patience 50 rounds was also utilized during training using MAE as the evaluation metric. This helped ensure good generalization and prevented overfitting because the model would stop training when the validation error no longer improved.

3.6.4 RANDOM FOREST

Random Forest (RF) is an ensemble learning tree-based algorithm widely recognized for its power, interpretability, and high predictive accuracy on regression and classification tasks (Breiman, 2001). RF operates by training numerous decision trees and estimating the average (in regression) or majority vote (in classification) of individual tree prediction. The ensemble method avoids overfitting and generalizes the model better, particularly in noisy or high-dimensional data.

Random Forest exploits two sources of randomness contributing to the performance: bootstrapping (data resampling with replacement for tree training) and random choice of the split features. These operations inject diversity among trees, which translates its high accuracy and robustness to the model. RF is not distribution assumption-dependent, so it can be used for a wide variety of problems from real life, such as data sets with severely skewed distributions, missing values, or outliers.

Random Forest is one of the key advantages of feature importance ranking, which provides a sense of what variables have the most impact for model output. This is very useful in fields such as retail forecasting, healthcare, and environmental modelling, where interpretability is as important as predictive accuracy. Random Forest has performed well in several forecasting problems, including product demand prediction, crop yield estimation, and customer churn modelling, and has been found to be versatile in various domains (Evers et al., 2018; Raizada & Saini, 2021; Liaw & Wiener, 2002).

3.6.4.1 MODELLING - RANDOM FOREST

As one of the weekly sale's predictive models for a football club's shop, a Random Forest Regressor from the scikit-learn library was utilized. Being an ensemble model, it is well suited for problems with regression and multiple variables, as it assigns the decisions from numerous decision trees to smooth out the outcome, hence avoiding overfitting and enhancing generalization (Couronné et al., 2018).

The model was trained on the training dataset with 100 trees and a fixed random seed for reproducibility. Following the fitting of the model, the relative importance of every input feature was tested to determine which variables made the biggest contribution to the predictions. The importance values were ordered, and the 20 most important features were visualized using a horizontal bar plot. This segmentation gave insightful insights on the major drivers of weekly sales performance.

3.6.4.2 HYPERPARAMETER TUNING – RANDOM FOREST

For optimizing the performance of the Random Forest model for predicting weekly sales in the football club's store, two-step hyperparameter optimization was used. To begin with, RandomizedSearchCV was used, whose hyperparameter space could be searched efficiently by sampling randomly from a defined grid. It contained 50 trials times 5-fold cross-validation, or 250 fits in total, and it was evaluating performance as negative mean squared error. The optimal parameters found in this phase were: n_estimators: 300; max_depth: None; min_samples_split: 2; min_samples_leaf: 1; max_features: None; bootstrap: True.

Although these findings gave a good foundation, additional optimization was implemented using Optuna, a sophisticated hyperparameter optimization system rooted in Bayesian search methods. Optuna tested extra combinations by repeatedly validating and training new Random Forest models on randomly chosen configurations, employing a holdout validation set to provide equitable evaluation. Each trial's performance was recorded with the MAE.

Upon running 50 trials, Optuna concluded the optimal set of hyperparameters to be: n_estimators: 330; max_depth: 59; min_samples_split: 2; min_samples_leaf: 2; max_features: 0.8897

This configuration resulted in a minimum validation MAE of about 38,908.79. A final model was then trained using the whole training set with these optimized parameters to yield a more powerful and robust forecasting model for weekly sales.

3.7 EVALUATION

This section presents an objective overview of the results obtained from the application of four predictive models to the dataset under study. The models evaluated are XGBoost, SARIMAX, LightGBM, and Random Forest. The goal of this section is to document the performance of each model based on predefined evaluation metrics, without interpreting or discussing the implications of the results — that analysis will be addressed in a later chapter. The results for each model are reported systematically below.

The performance was also judged using three of the most popular performance metrics used in predictive modelling, namely RMSE, MAE, and the Coefficient of Determination (R^2)
RMSE is a standard metric, which computes the square root of the mean of the squared difference between actual and predicted values. RMSE gives more weightage to big errors due to the squaring and is more outlier sensitive than MAE (Hodson, 2022).. For a sample of n observations y (yi, i = 1,2,..., n) and n corresponding model predictions yˆ, the RMSE is (Hodson, 2022):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

RMSE is useful in evaluating models, in a lot of scenarios, especially when large errors are particularly undesirable (Wang & Lu, 2018).

MAE calculates the average of the absolute differences between predicted and actual values. Unlike RMSE, it treats all errors equally and is less sensitive to outliers, making it easier to interpret (Willmott & Matsuura, 2005). For a sample of n observations y (yi, i = 1,2,..., n) and n corresponding model predictions yˆ, the MAE is (Hodson, 2022):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

Adjusted R^2 represents the proportion of the variance in the target variable that is explained by the model. It refines the R^2 measure by accounting for the number of predictors in the model and the sample size, providing an overall indication of goodness-of-fit, with values closer to 1 indicating better predictive performance (Miles, 2014). However, its key advantage lies in its ability to provide a more reliable measure of goodness-of-fit when comparing models that include a different number of independent variables (Bar-Gera, 2017). While R^2 will always increase or stay the same with the addition of new predictors, regardless of their significance, adjusted R^2 penalizes the inclusion of unnecessary terms, and may decrease if a predictor does not improve the model significantly or beyond what would be expected by chance (Miles, 2014).

$$Adj.R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

These metrics are considered appropriate for evaluating predictive models of both regression and time series nature because they capture different aspects of error and model performance. RMSE and MAE provide direct measures of prediction accuracy in the same units as the target variable, while Adjusted R^2 offers a normalized indicator of explanatory power, making comparisons between models more interpretable (Zhang & Lu, 2020).

Metrics were calculated on the validation set and, finally, on the test set (kept untouched until the final evaluation) to obtain a realistic estimate of the model's performance on future data. In addition to statistical evaluation, model assessment was also considered from a business perspective. This involved analysing whether the achieved accuracy was sufficient to support inventory management and planning decisions, and whether the forecasts made sense within the context of business knowledge. Comparison among different candidate models allowed for the selection of the approach that offered the best balance between statistical accuracy and practical utility.

4. RESULTS AND DISCUSSION

This chapter presents the evaluation of the predictive models developed in the context of this study. The primary objective is to assess the performance of each model in forecasting weekly sales (VAL_SEM_IVA) and to identify the most suitable approach for deployment. Four models were tested: XGBoost, SARIMAX, LightGBM, and Random Forest. Each model is first evaluated individually using standard predictive performance metrics and then compared against the others in a structured analysis.

The evaluation relies on three widely accepted regression metrics RMSE, MAE, and the Coefficient of Determination (R^2). These metrics were chosen for their complementary perspectives on model accuracy.

Beyond statistical accuracy, model behaviour over time is also considered through forecast plots and residual analysis. These visual tools support the interpretation of how each model performs across different sales patterns, such as peaks, drops, and stable periods. In addition, business interpretability and practical usefulness are considered, as the selected model is intended for real-world application in sales planning.

The chapter concludes with a comparative synthesis of all models, leading to the selection of the most effective one based on its overall performance and suitability for deployment.

4.1 RESULTS ANALYSIS

4.1.1 XGBOOST

The XGBoost model, after the final optimization with the Optuna library, gave the following performance metrics, RMSE of 84, a MAE of 46010.88 and an adjusted R^2 of 0.846.

These statistics outline the model's test set prediction performance, with an adjusted R^2 of approximately 0.846 suggesting the model explains a very large percentage of the variance in the target variable (VAL_SEM_IVA), which is the club's sales per week excluding VAT. RMSE and MAE document the mean magnitude of the prediction errors, in euros.

As an example of MAE, the MAE value of 45,357.88 € means that the average deviation from the model's forecast of weekly sales and actual sales is approximately 45,000 €, while the RMSE implies that bigger errors may occasionally occur. Despite these discrepancies, the high R^2 value suggests that the model accurately depicts the overall tendency of sales.

The figure 11 compares the actual and predicted values of VAL_SEM_IVA over time- It shows that the model closely follows the rise and fall in weekly sales, with peaks and troughs. Although some differences exist between the size of some peaks (e.g., at weeks 0 and 21), the model can well estimate the general shape and seasonal pattern of the time series. This visual stability reinforces the quantitative measures and shows that XGBoost can learn temporal structures within the data, especially when lag features are used.

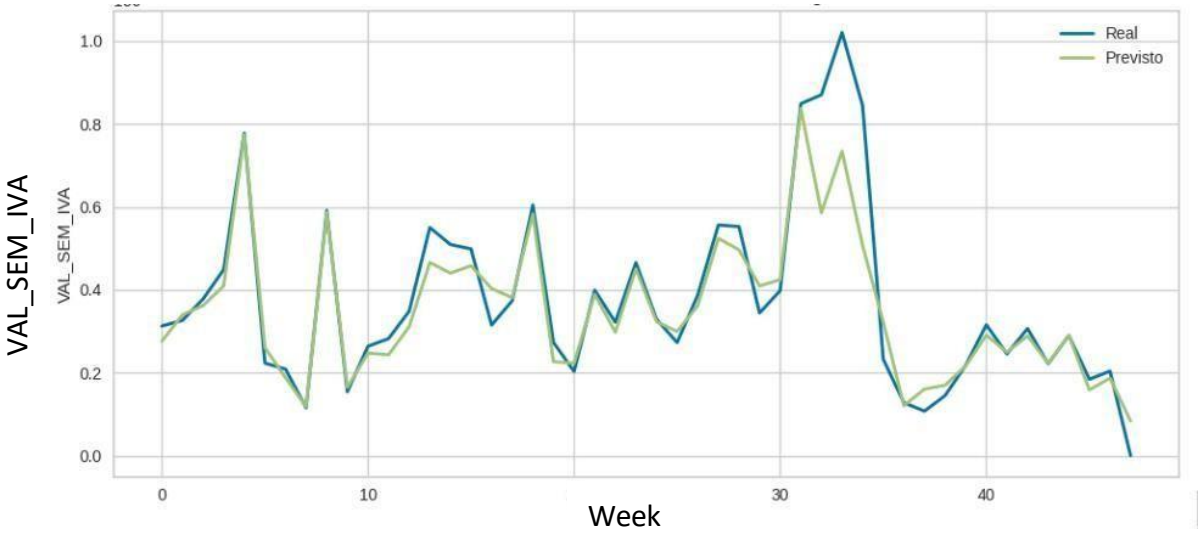

Figure 11 - Actual VS Predicted values of VAL_SEM_IVA over time – XGBoost

The plot of residuals (Real – Predicted) also illustrates the behaviour of the model. The residuals are quite symmetrically distributed around zero across the weeks, with no systematic over- or underestimation bias. While some of the larger discrepancies do occur (specifically around weeks 0, 10, and 33), most of the residuals remain in bounds, and that suggests consistency on the part of the model. The absence of strong pattern or trend on the part of the residuals also suggests that the model did effectively capture underlying structure of the data.

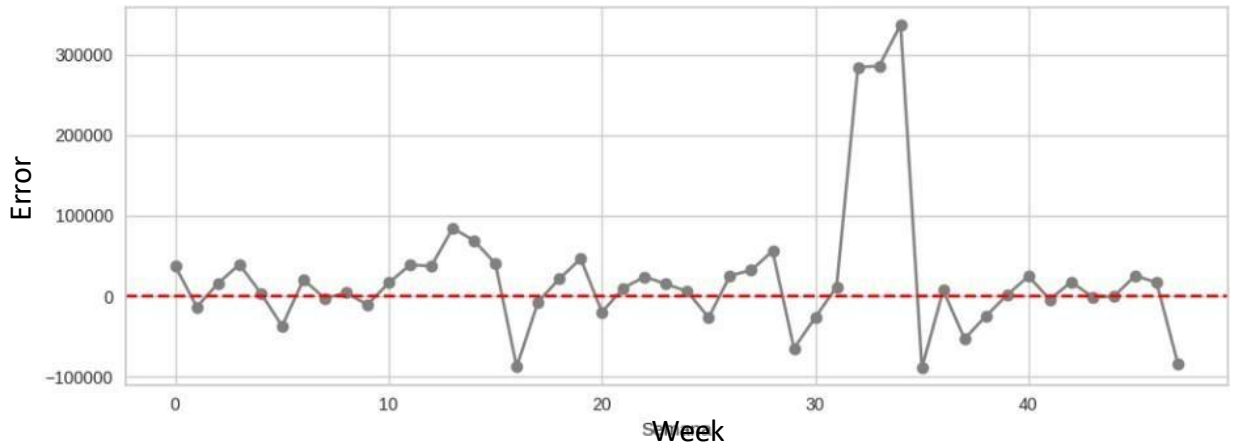

Figure 12 – Plot of Residuals over time -XGBoost

4.1.2 SARIMAX

SARIMAX model was selected after performing a grid search across various seasonal and non-seasonal configurations manually. The final model had both autoregressive and seasonal terms along with a few exogenous variables. Its performance on the test set is based on a RMSE of 399,905.31, a MAE of 282,618.99 and an adjusted R^2 of -5.200

These results indicate that the model's predictions deviated considerably from real weekly sales (VAL_SEM_IVA) with an average absolute error of over 280,000 € and some errors even larger in size. The negative adjusted R^2 value signals that the model fails to beat a naively simple mean-based prediction, i.e., that it fails to capture the underlying trend in the test data.

The test-period predicted vs. actual values plot (figure 13) of the plot shows that the SARIMAX model struggled to follow the actual sales pattern during the test period. The model correctly identifies the periods when the fluctuations take place but overestimates the amplitude of some of the peaks (e.g., during the first weeks of 2022) and gets lost or falls behind the actual downward sales movement toward the latter part of the forecast horizon.

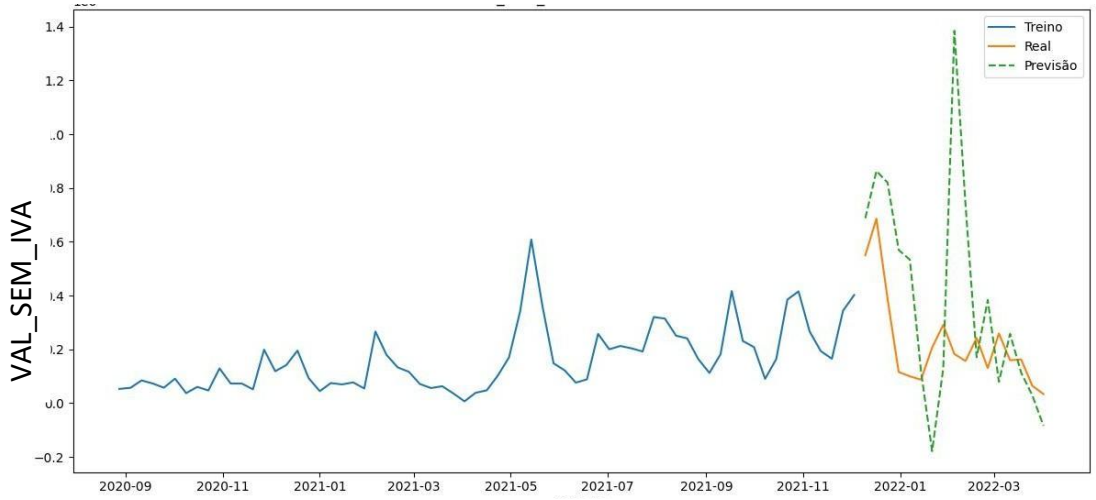

Figure 13 - Actual VS Predicted values of VAL_SEM_IVA over time – SARIMAX

The residuals plot also confirms this phenomenon. The residuals (Real – Predicted) are scattered all over the place and not symmetrically located around zero, some of them even higher than –1.2 million euros. This is a sign of the absence of stability and consistency in the predictability power of the model, perhaps due to the limited number of test observations, large fluctuation in weekly sales, or poor seasonal structure in the adopted configuration.

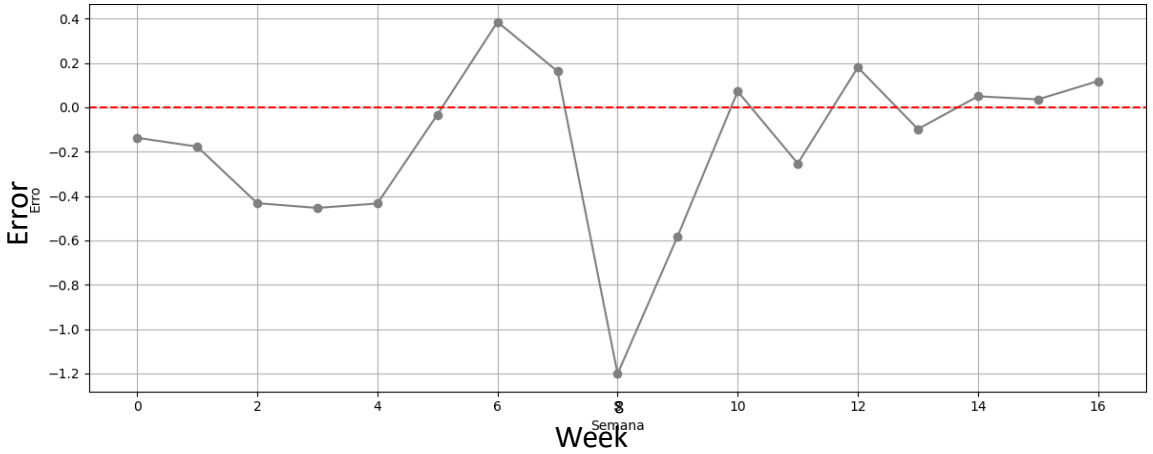

Figure 14 – Plot of Residuals over time -XGBoost

Overall, while theoretically SARIMAX is powerful in the modelling of time series with exogenous regressors, the model did not handle the present forecasting task well enough against the other approaches being tested.

4.1.3 LIGHTGBM

The LightGBM model was trained with the same target variable (VAL_SEM_IVA) and exogenous regressors. Its performance on the test set is as follows MAE of 34,783.88, a RMSE of 46,649.52 and an adjusted R^2 of 0.7680

These results demonstrate that the LightGBM model performed quite well in weekly sales prediction. That R^2 of 0.7704 demonstrates that the model explained approximately 77% of VAL_SEM_IVA variance. The MAE of around 34,784 € signifies that predictions were off actual weekly sales by that amount on average, and the RMSE of 46,649.52 € testifies that there are some bigger errors but that they are kept fairly in line.

The figure 15 shows predicted against actual values, demonstrates that the LightGBM model captures the overall trend of weekly sales extremely well. Peaks and troughs (particularly at weeks 6 and 10) are mostly found in the predictions, though there are some amplitude discrepancies—particularly the peak at week 7, when the model underestimates the amplitude.

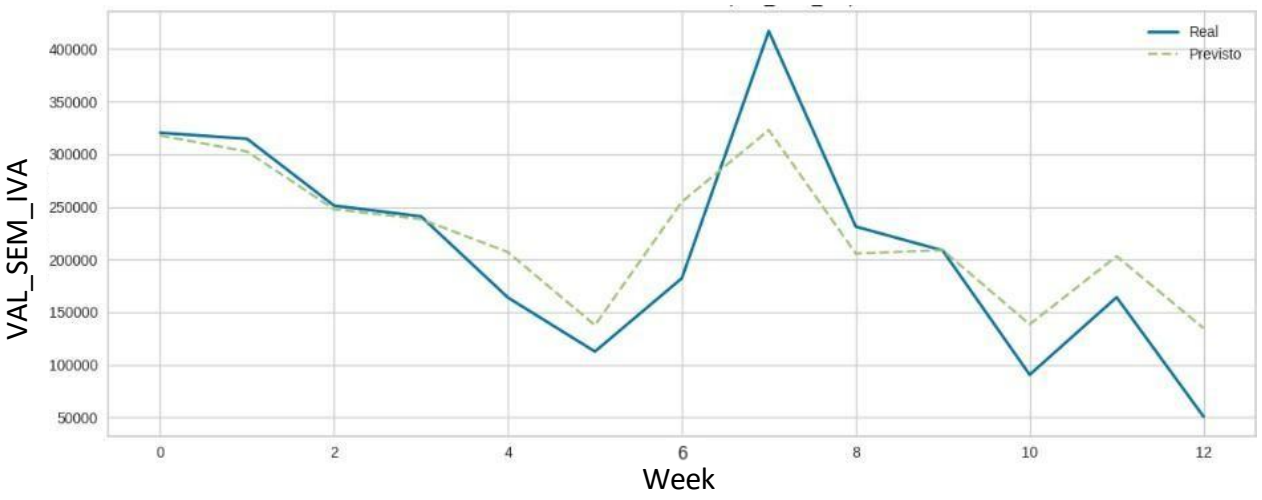

Figure 15 - Actual VS Predicted values of VAL_SEM_IVA over time – LightGBM

The plot of residuals (Real – Predicted) (figure 16) shows most prediction errors are very small and scatter around zero. Though there is one huge positive residual in week 7 (representing underestimation of one sales peak), most residuals vary between ±50,000 €, showing consistent behaviour and no prominent systematic distortion.

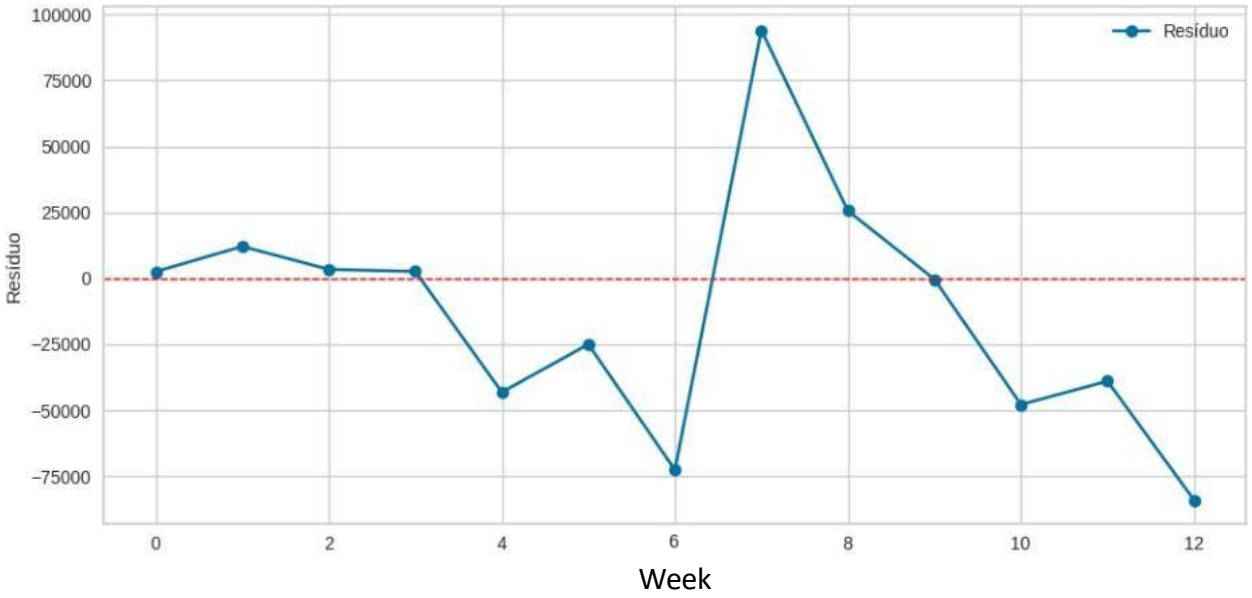

Figure 16 – Plot of Residuals over time - LightGBM

Generally, the LightGBM model works very well, working well at rendering temporal variability and yielding good predictive reliability on the weekly sales predictions.

4.1.4 RANDOM FOREST

The test set performance of the model is based on a RMSE of 108,754.56, a MAE of 56,329.96 and an adjusted R^2 of 0.7450.

These results indicate a mid-strength performance. The adjusted 0.7450 R^2 shows that the model accounts for roughly 75% of the variation in weekly sales. The MAE of slightly more than 56,000€ implies that on average, predictions diverge from actual sales by that amount. The higher RMSE, in comparison to MAE, shows the presence of larger individual errors, especially for weeks of sharp spikes.

Forecast vs. actual plot shows that although the model is tracking the overall increasing trend along the forecast horizon, it is suddenly underestimating the steep peak in sales experienced in week 6. While the forecasted curve is smooth and even, the actual values rise at a faster rate, peaking higher than the model is unable to track in size.

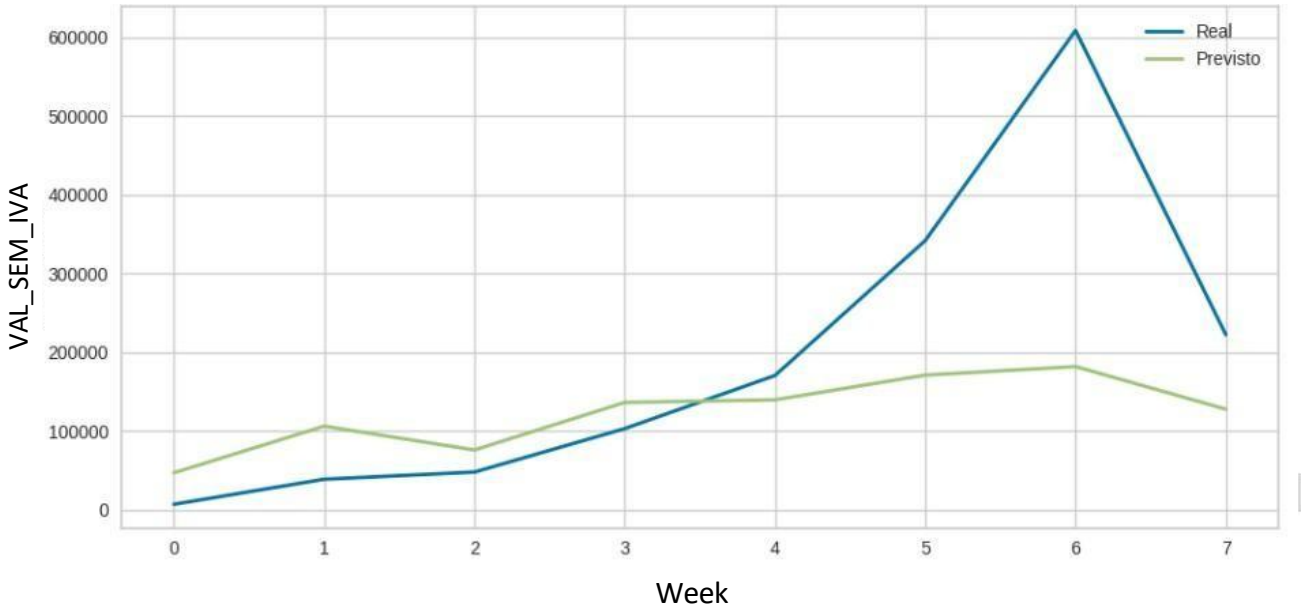

Figure 17 – Actual VS Predicted values of VAL_SEM_IVA over time – Random Forest

The residual plot (figure 18) substantiates this, with a positive error of the high magnitude in week 6, where the model had underpredicted more than 400,000 €. Otherwise, the errors are comparatively small and comparatively constant, though they do exhibit a minor trend toward underprediction in the following weeks.

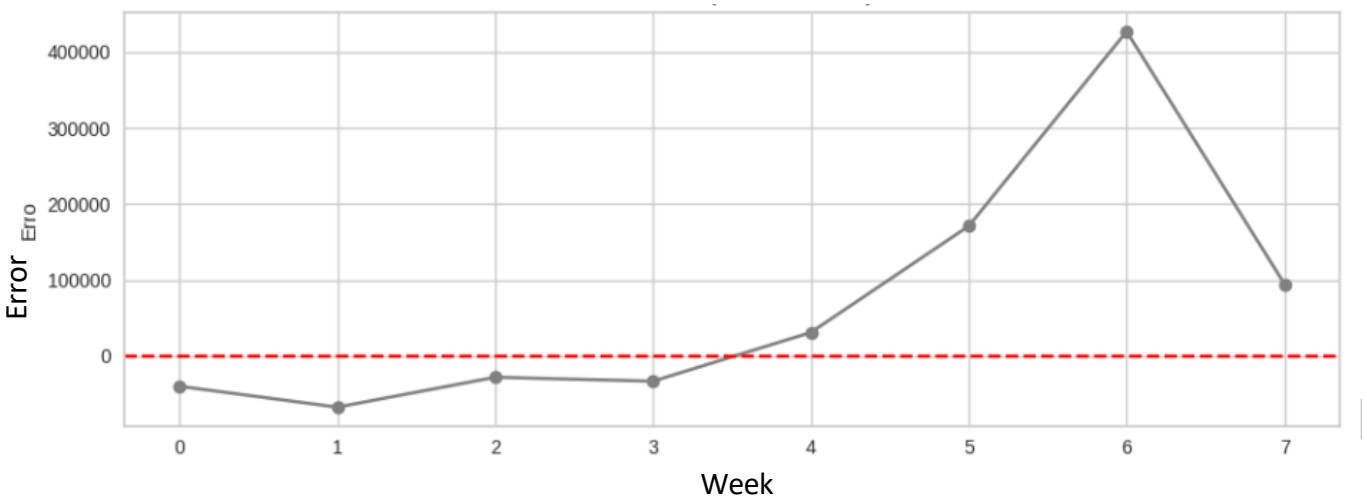

Figure 18 – Plot of Residuals over time – Random Forest

Overall, the Random Forest model demonstrates decent predictive performance but is prone to over smoothing sudden changes because of its ensemble structure and lack of temporal dynamics. It performs optimally in capturing gradual trends but fails with extremely volatile or random weekly spikes in sales.

The project successfully achieved the initial business goals by providing accurate predictive models that support inventory management and decision-making, demonstrating adequate statistical performance and practical utility in the business context. After considering potential

actions, the recommended decision is to proceed with model deployment while setting up monitoring procedures. This choice balances accuracy, business value, and resource availability.

4.2 Discussion of Results

After the analysis of every model separately, a comparison was made to ascertain their relative performance at forecasting weekly sales (VAL_SEM_IVA). The comparison considers not only predictive accuracy as measured by RMSE, MAE, and R^2, but also each model's behaviour across time, residual plots, and business decision-making practical considerations. The results are in the table 4.

Table 4 – Comparative Performance of the models.

Model	RMSE (€)	MAE (€)	Adjusted R^2
XGBoost	84,221.57	46,010.88	0.846
SARIMAX	399,905.31	282,618.99	-5.200
LightGBM	46,649.52	34,783.88	0.7680
Random Forest	108,754.56	56,329.96	0.7450

Though LightGBM performed the lowest error rates in both RMSE and MAE, XGBoost performed the highest adjusted R^2 score, explaining nearly 85% of weekly sales variance. It suggests that XGBoost generalizes better and captures bigger trends more effectively, even though it is slightly less accurate for point-to-point estimation. Random Forest performed moderately but with higher errors in general. SARIMAX, however, yielded very unstable results with very high error and negative R^2, indicating poor model fit and minimal forecasting ability.

To more readily observe practical differences between these models, the figure 19 offers a graphical comparison of RMSE, MAE, and adjusted R^2 for all four methods. This helps to highlight trade-offs between explanatory power and raw accuracy.

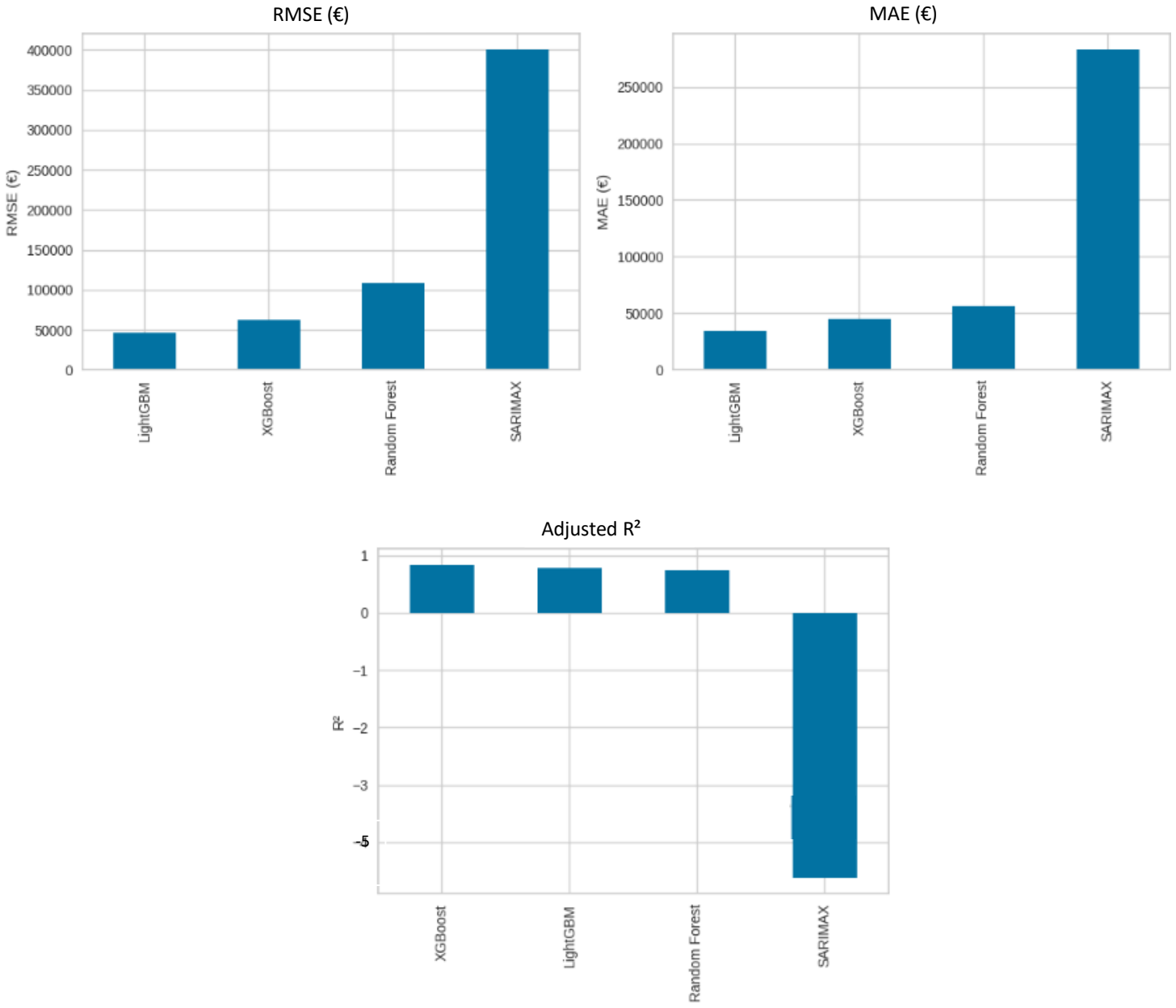

Figure 19 - Comparison of RMSE, MAE, and adjusted R^2 for all four methods (bar charts)

Beyond metric values, temporal prediction behaviour also plays a crucial role. When comparing predicted vs. actual weekly sales over the test period, LightGBM and XGBoost were able to capture most of the upward and downward movements in sales. However, LightGBM occasionally underestimated sharp increases, particularly around week 7. XGBoost, although not perfect, followed the trend more consistently and aligned well with seasonal effects. Random Forest often smoothed out volatility, failing to account for sudden peaks. SARIMAX frequently either lagged actual changes or overestimated values unpredictably. An overall comparison between the models' performances can be seen in the figure 20.

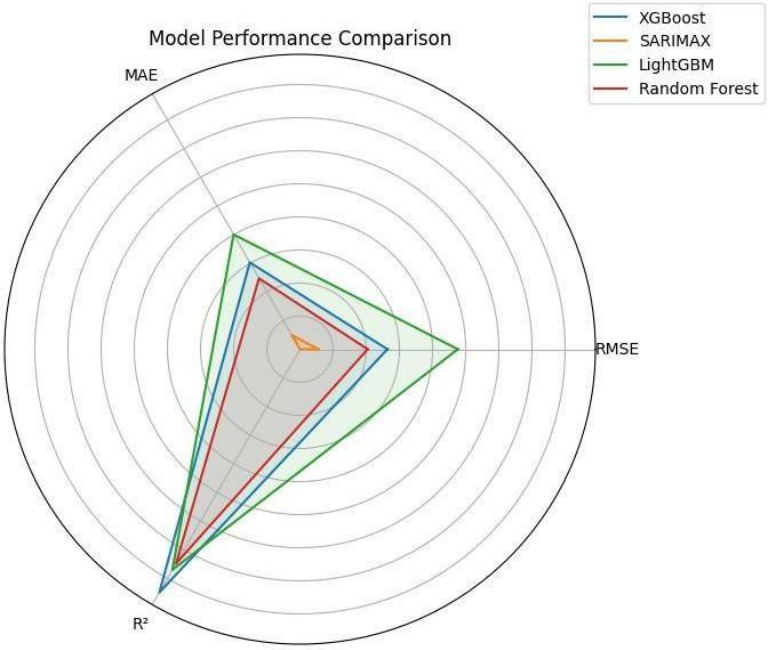

Figure 20 - Comparison of RMSE, MAE, and adjusted R^2 for all four methods (Spider Chart)

To better understand how the selected model (XGBoost) arrives at its predictions, a SHAP analysis was performed. SHAP values provide a unified measure of feature importance by quantifying the contribution of each input variable to the final prediction for each instance. Unlike traditional feature importance metrics that only show average effect, SHAP offers both global interpretability (which features matter most overall) and local interpretability (why a specific prediction was made) (Lundberg et al., 2018).

The figure 21 shows the global summary plot of SHAP values: each dot represents a prediction for a specific week, and its position along the x-axis indicates how much that feature increased or decreased the predicted sales value. Colour represents the actual value of the feature for that prediction (e.g., high or low temperature, presence or absence of a campaign, etc.).

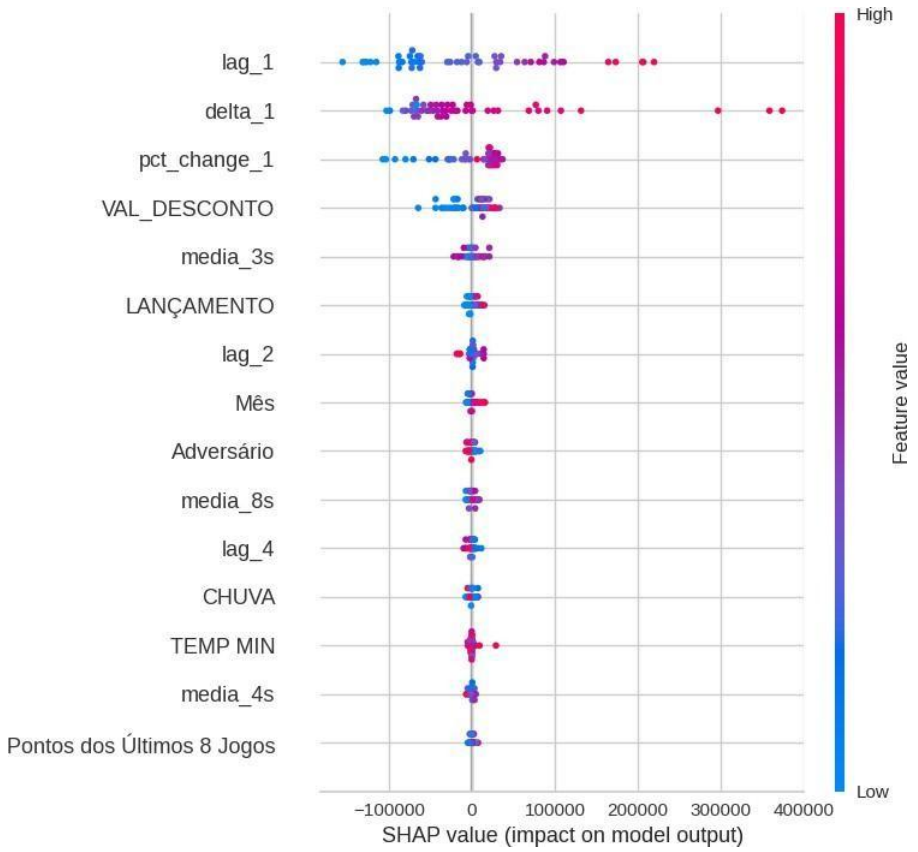

Figure 21 – SHAP Summary Plot (Feature Impact by Value)

The variables are ordered by their mean contribution towards the model output for all test set instances, thereby making it easy to identify the most contributing factors. lag_1 seems to be the strongest predictor, and this indicates that the previous week's value of sales is the strongest driver of the model's prediction. This finding aligns with the temporal nature of the task, where recent sales are naturally strong indicators of near-future performance.

Following closely, delta_1 and pct_change_1, which represent recent changes and percentage variation in sales, also show considerable influence. These variables suggest that short-term sales dynamics and momentum are captured effectively by the model and play a crucial role in shaping its forecasts. The variable VAL_DESCONTO, which reflects the total value of discounts applied in each week, further underscores the importance of promotional activity. Its high importance score confirms that sales campaigns are key explanatory factors for fluctuations in weekly revenue.

Other variables such as media_3s and LANÇAMENTO also contribute meaningfully to the model’s output. The presence of recent product launches appears to have a measurable effect on predicted sales, which is consistent with the expected boost in consumer activity following new releases. The colour gradient used in the plot, which encodes the original feature values, adds another interpretative layer. For instance, higher values of lag_1 (in pink) are likely to be correlated with higher values of SHAP, so weeks with stronger prior sales are likely to lead to

higher sales prediction. This pattern is an evident sign that the model is showing statistically as well as contextually significant behaviour.

By and large, the SHAP summary plot is not only highlighting the features most actively engaged but also how exactly these variables contribute to engaging with the prediction process. This level of transparency enhances the explainability of the XGBoost model and lends credibility to its deployment within a live forecasting context.

5. CONCLUSION

This study was focused on developing a stable predictive model that can forecast weekly sales revenue (VAL_SEM_IVA) of a sports club by combining machine learning and time series modelling methods. Based on a step-by-step feature engineering process, training of models, evaluation, and exploration of interpretability, four models were compared: XGBoost, LightGBM, Random Forest, and SARIMAX.

The testing stage revealed substantial differences in model performance. As much as LightGBM had the lowest RMSE and MAE scores, reflecting improved performance in raw predictive accuracy, XGBoost was the one with the highest R^2 score and managed to explain approximately 85% of weekly sales variance. This reflects that XGBoost provides better generalization, particularly in explaining overall sales trends. Random Forest gave average results but with increased error rates and a smoothing effect on large fluctuations. SARIMAX, despite its traditional use in time series modelling, was not great in this case with high error and instability in prediction.

In addition to numerical performance, the models were also compared in temporal behaviour and residual pattern. XGBoost and LightGBM performed best in detecting both incremental and break changes in sales, although LightGBM underestimated sudden sales boosts in certain weeks. XGBoost was more regular and representative of observed seasonality and therefore more appropriate for business-related forecasting. These findings were confirmed by visual observations of actual vs. predicted and residual plots, which disclosed model behaviour that raw measures would not be able to fully account for.

By and large, the project was a success in identifying XGBoost as the most suitable model for forecasting weekly sales in this scenario. It is robust in predictive performance with good generalization, handles feature interactions well, and possesses transparency with interpretability tools. This makes XGBoost not just a high-performance algorithm but also a sound decision-support tool for weekly revenue prediction. Future extensions could involve continuing the integration of other external variables (i.e., economic indicators, competitor actions) or creating real-time pipelines for dynamically refreshing predictions.

5.1 DEPLOYMENT

The deployment process consisted of integrating the trained XGBoost model into a simplified forecasting pipeline designed to operate on a weekly basis. This pipeline includes preprocessing steps (such as generation of lag variables and normalization of exogenous inputs), model inference, and output formatting to facilitate business interpretation. The objective is to provide the club with reliable weekly sales forecasts (VAL_SEM_IVA) to support planning in areas such as inventory management, campaign scheduling, and staffing.

Additionally, safeguards were included to ensure robustness in operational use, these included:

Validation of input data: Automatic checks to identify missing or anomalous values before generating predictions.

Forecast horizon limitation: Restricting predictions to a maximum of 6 weeks ahead, consistent with the training setup and minimizing extrapolation risks.

Retraining schedule: Implementation of a retraining scheme monthly, allowing the model to incorporate current sales and external factors to enable flexibility with changing dynamics.

While the model is running in batch form now, its modular structure means it can in the future be integrated within a live dashboard or API. This will make it more accessible for organizational users without technical expertise.

The deployment stage not only deployed the top-performing model that was found in testing but also ensured that the prediction solution is aligned with the real needs of the business, production stable, and scalable for future developments.

5.2 LIMITATIONS

While this project successfully isolated and utilized XGBoost as the most suitable model to forecast weekly sales (VAL_SEM_IVA), there are real limitations that suggest future research opportunities.

Many variables were not cross-sectional across multiple sales datasets, while others were empty. This missed an opportunity to explore more fields and obtain more useful information to generate good performance for the applied predictive models.

Deployment was in batch mode on a weekly schedule reducing responsiveness to rapidly changing conditions, like promotions or random bursts of demand, that would best respond to real-time or near-real-time forecasting.

Interpretability is also a pragmatic limitation, although XGBoost does provide such as feature importance and SHAP values, it can still prove complicated for non-technical individuals to completely understand why the forecast changes week-to-week, potentially limiting its application in business decision-making on a frequent basis.

5.3 FUTURE RESEARCH

For future research, a larger feature set that integrates more richly structured external data sources could improve predictiveness and stability under market changes. Exploration of hybrid modelling methods that combine machine learning with advanced time-series architecture could lead to better performance on change in trends and seasonality patterns.

Extension of the forecast horizon with mitigation of uncertainty, for example through probabilistic forecasting methods, could add value for use in strategic planning.

Moving the deployment pipeline to real-time predictions instead of batches would increase agility in responding to a shifting market, and automated retraining and monitoring systems can maintain model performance as business conditions shift. Finally, having more accessible and user-friendly interpretability tools, e.g., dashboards or natural-language descriptions, available would make projections more explainable and transparent to non-technical stakeholders and allow for easier support by them in business planning through increased trust.

6. BIBLIOGRAPHICAL REFERENCES

Ahn, T., Suh, Y., Lee, J., & Pedersen, P. (2012). Sport Fans and Their Teams' Redesigned Logos: An Examination of the Moderating Effect of Team Identification on Attitude and Purchase Intention of Team-Logoed Merchandise. Journal of Sport Management, 27, 11-23. https://doi.org/10.1123/JSM.27.1.11

Alim, M., Ye, G., Guan, P., Huang, D., Zhou, B., & Wu, W. (2020). Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. https://doi.org/10.1136/bmjopen-2020-039676

Aljohani, A. (2023). Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility. *Sustainability*, *15*(20), 15088. https://doi.org/10.3390/su152015088

Ampountolas, A. (2021). Modeling and Forecasting Daily Hotel Demand: A Comparison Based on SARIMAX, Neural Networks, and GARCH Models. Forecasting, 3(3), 580-595. https://doi.org/10.3390/forecast3030037

Appelqvist, P., Babongo, F., Chavez-Demoulin, V., Hameri, A., & Niemi, T. (2016). Weather and supply chain performance in sport goods distribution. International Journal of Retail & Distribution Management, 44(2), 178–202. https://doi.org/10.1108/ijrdm-08-2015-0113

Ascenção, J. (2023). Forecasting demand in the pharmaceutical industry using machine learning (Master's thesis, Universidade NOVA de Lisboa (Portugal). https://run.unl.pt/handle/10362/159901

Assad, S., Clark, R., Ershov, D., & Xu, L. (2023). Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. Journal of Political Economy, 132, 723 - 771. https://doi.org/10.2139/ssrn.3682021

Atanda, O., Adebiyi, M., Adewumi, D., Abiodun, M., Awodoye, O., & Adebiyi, A. (2024). Intelligent Sales Forecasting System Using Arima, Sarima, and Xgboost Models. 1–8. https://doi.org/10.1109/seb4sdg60871.2024.10629780

Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, M., & Marquis, P. (2022). Trading Complexity for Sparsity in Random Forest Explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 36(5), 5461-5469. https://doi.org/10.1609/aaai.v36i5.20484

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. In International Journal of Forecasting , 35(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Bar-Gera, H. (2017). The Target Parameter of Adjusted R-Squared in Fixed-Design Experiments. *The American Statistician*, *71*(2), 112–119. https://doi.org/10.1080/00031305.2016.1200489

Bastos, A. (2024). Machine Learning in Digital Retail: Demand Forecasting for Inventory Management in a Sportswear Company (Master's thesis, Universidade NOVA de Lisboa (Portugal). https://run.unl.pt/handle/10362/175290

Ben-Bouallegue, Z., Clare, M., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J., Lang, S., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., & Pappenberger, F. (2023). The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. Bulletin of the American Meteorological Society. 105(6), E864-E883. https://doi.org/10.1175/bams-d-23-0162.1

Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review 54, 1937–1967 (2021). https://doi.org/10.1007/s10462-020-09896-5

Breitbarth, T., & Harris, P. (2008). The role of corporate social responsibility in the football business: towards the development of a conceptual model. European Sport Management Quarterly, 8(2), 179–206. https://doi.org/10.1080/16184740802024484

Breskvar, M., Kocev, D., & Džeroski, S. (2018). Ensembles for multi-target regression with random output selections. Machine Learning, 107(11), 1673–1709. https://doi.org/10.1007/s10994-018-5744-y

Bruckhaus, T. (2007). The Business Impact of Predictive Analytics. In X. Zhu & I. Davidson (Eds.), Knowledge Discovery and Data Mining: Challenges and Realities (pp. 114-138). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-59904-252-7.ch007

Budholiya, K., Shrivastava, K., & Sharma, V. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University - Computer and Information Sciences, 34(7), 4514–4523. https://doi.org/10.1016/j.jksuci.2020.10.013

Bulgakova, S., & Zosimov, V. (2024). Modeling the Impact of External Factors on E-Commerce Consumer Behavior. 2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT). IEEE. https://doi.org/10.1109/CSIT65290.2024.10982577

Bunker, R., & Susnjak, T. (2019). The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review. Journal of Artificial Intelligence Research, 73, 1285-1322. https://doi.org/10.1613/jair.1.13509

Catto, L., & Dowdy, A. (2021). Understanding compound hazards from a weather system

perspective. Weather and Climate Extremes, 32, 100313. https://doi.org/10.1016/j.wace.2021.100313

Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Chatfield, C. (2001). Time-series forecasting. Chapman & Hall/CRC. https://doi.org/10.1201/9781420036206

Chen, Q. (2023). Enterprise marketing strategy using big data mining technology combined with XGBoost model in the new economic era. PLOS ONE, 18(6): e0285506. https://doi.org/10.1371/journal.pone.0285506

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785 – 794. https://doi.org/10.1145/2939672.2939785

Choi, T., Hui, C., & Yu, Y. (2011). Intelligent time series fast forecasting for fashion sales: A research agenda. International Conference on Machine Learning and Cybernetics, 1010 - 1014. https://doi.org/10.1109/icmlc.2011.6016870

Ciaburro, G., & Iannace, G. (2021). Machine Learning-Based Algorithms to Knowledge Extraction from Time Series Data: A Review. Data, 6, 55. https://doi.org/10.3390/data6060055

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 19(1). https://doi.org/10.1186/s12859-018-2264-5

Davagdorj, K., Pham, H., Theera-Umpon, N., & Ryu, K. H. (2020). XGBOOST-Based Framework for Smoking-Induced Noncommunicable Disease Prediction. International Journal of Environmental Research and Public Health, 17(18), 6513. https://doi.org/10.3390/ijerph17186513

Benitez, I., Ibañez, J., Lumabad, C., Cañete, J., & Principe, J. (2023). Day-Ahead hourly solar photovoltaic output forecasting using SARIMAX, Long Short-Term memory, and Extreme gradient boosting: Case of the Philippines. Energies, 16(23), 7823. https://doi.org/10.3390/en16237823

Deng, T., Zhao, Y., Wang, S., & Yu, H. (2021). Sales Forecasting Based on LightGBM. IEEE Xplore. https://doi.org/10.1109/ICCECE51280.2021.9342445

Eboigbe, E., Farayola, O., Olatoye, F., Nnabugwu, O., & Daraojimba, C. (2023). BUSINESS INTELLIGENCE TRANSFORMATION THROUGH AI AND DATA ANALYTICS. Engineering Science & Technology Journal. https://doi.org/10.51594/estj.v4i5.616.

Elhady, A., Shohieb, S., Tarek, Z., Abdelhamid, A., Ibrahim, A., Shams, M., & Elshewey, A.

(2022). A Novel WD-SARIMAX Model for Temperature Forecasting Using Daily Delhi Climate Dataset. Sustainability. https://doi.org/10.3390/su15010757.

Evers, M., Tavasszy, L., Van-Duin, R., Schott, D., & Gorte, F. (2018). Demand forecast models for online supermarkets. In E-groceries, digitalization and sustainability: Which governance, planning and regulation mix do our cities need? (pp. 1). Molde University.

Fotache, M., Cojocariu, I., & Bertea, A. (2021). High-Level Machine Learning Framework for Sports Events Ticket Sales Prediction, 55-60. https://doi.org/10.1145/3472410.3472426

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., & Wolff, E. (2018). Very high resolution Object-Based Land Use–Land cover urban classification using extreme gradient boosting. IEEE Geoscience and Remote Sensing Letters, 15(4), 607–611. https://doi.org/10.1109/lgrs.2018.2803259

Gómez, J., Schröer, C., & Kruse, F. (2020). A Systematic Literature Review on Applying CRISP-DM Process Model. https://doi.org/10.1016/J.PROCS.2021.01.199.

Gregório, B. (2021). The impact of marketing on the value of a brand of a football club http://hdl.handle.net/10400.26/38585

Groebner, D. F. (1990). Solving the inventory problem for the sale of seasonal merchandise - Journal of Small Business Management, 28(3) https://www.proquest.com/openview/4ec54dfe0d96262296ba62dfeb26ab0a/1?cbl=49244&pq-origsite=gscholar

Habenstein, D., Kirchhoff, K., & Schlesinger, T. (2020). Club fan shop or not? A conjoint analysis of online jersey purchase behavior. Sport Business and Management an International Journal, 11(1), 54–71. https://doi.org/10.1108/sbm-10-2019-0102

Hedlund, D. (2014). Creating value through membership and participation in sport fan consumption communities. European Sport Management Quarterly, 14, 50 - 71. https://doi.org/10.1080/16184742.2013.865775.

Henriques, P. (2024). Shipping Volume Forecasting in an International Lifestyle Company: Comparing Time Series Forecasting techniques. http://hdl.handle.net/10362/174745

Hodson, O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development, 15(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Huang, F., & Boutros, C. (2016). The parameter sensitivity of random forests. BMC Bioinformatics, 17(1). https://doi.org/10.1186/s12859-016-1228-x

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019) Automated Machine Learning: Methods, Systems, Challenges; Springer Nature:Berlin/Heidelberg, Germany p. 219. https://doi.org/10.1007/978-3-030-05318-5

Irfani, A., Supriyanto., & Pradipto, G. (2024). Optimization of Raw Material Inventory using Always Better Control (ABC) Analysis and Economic Order Quantity (EOQ) Method Approach in the Warehouse of a Bolt Manufacturing Factory in Indonesia. International Journal of Innovative Science and Research Technology (IJISRT). https://doi.org/10.38124/ijisrt%2Fijisrt24jul654

Jackson, I., Ivanov, D., Dolgui, A., & Namdar, J. (2024). Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation. International Journal of Production Research, 62(17), 6120–6145. https://doi.org/10.1080/00207543.2024.2309309

Joel, T., & Oguanobi, U. (2024). Data-driven strategies for business expansion: Utilizing predictive analytics for enhanced profitability and opportunity identification. International Journal of Frontiers in Engineering and Technology Research, 6(02), 071-081. https://doi.org/10.53294/ijfetr.2024.6.2.0035

Karabadji, I., Korba, A., Assi, A., Seridi, H., Aridhi, S., & Dhifli, W. (2023). Accuracy and diversity-aware multi-objective approach for random forest construction. Expert Systems With Applications, 225, 120138. https://doi.org/10.1016/j.eswa.2023.120138

Khalid, S., Goldenberg, M., Grantcharov, T., Taati, B., & Rudzicz, F. (2020). Evaluation of deep learning models for identifying surgical actions and measuring performance. JAMA Network Open, 3(3), e201664. https://doi.org/10.1001/jamanetworkopen.2020.1664

Komolafe, A., Aderotoye, I., Abiona, O., Adewusi, A., Obijuru, A., Modupe, O., & Oyeniran, O. (2024). HARNESSING BUSINESS ANALYTICS FOR GAINING COMPETITIVE ADVANTAGE IN EMERGING MARKETS: A SYSTEMATIC REVIEW OF APPROACHES AND OUTCOMES. International Journal of Management & Entrepreneurship Research. https://doi.org/10.51594/ijmer.v6i3.939.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2021). On the selection of forecasting accuracy measures. Journal of the Operational Research Society, 73(5), 937–954. https://doi.org/10.1080/01605682.2021.1892464

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences, 22(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kunft, A., Katsifodimos, A., Schelter, S., Breb, S., Rabl, T., & Markl, V. (2019). An intermediate representation for optimizing machine learning pipelines. Proceedings of the VLDB Endowment, 12(11), 1553–1567. https://doi.org/10.14778/3342263.3342633

Li, H., Ji, Z., Liu, B., Li, M., & Luo, J. (2024). Intelligent Productivity Transformation: Corporate Market Demand Forecasting With the Aid of an AI Virtual Assistant. Journal of Organizational and End User Computing, 36, 1-27. https://doi.org/10.4018/joeuc.336284

Lipovetsky, S. (2022). Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. Technometrics, 64, 145 - 148. https://doi.org/10.1080/00401706.2021.2020521.

Lundberg, M., Erion, G., & Lee, S. (2018). Consistent individualized feature attribution for tree ensembles. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1802.03888

Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., & Wang, Z. (2021). XGBoost-based method for flash flood risk assessment. Journal of Hydrology, 598, 126382. https://doi.org/10.1016/j.jhydrol.2021.126382

Matsuki, K., Kuperman, V., & Van-Dyke, A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. Scientific Studies of Reading, 20(1), 20–33. https://doi.org/10.1080/10888438.2015.1107073

Miles, J. (2014). R squared, adjusted R squared. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06627

Mitrea, A., Lee, M., & Wu, Z. (2009). A Comparison between Neural Networks and Traditional Forecasting Methods: A Case Study. International Journal of Engineering Business Management, 1, 11. https://doi.org/10.5772/6777

Murugan, S., Selvi, P., Dheeksha, S., Deepalakshmi, R., & Mithun, S. (2023). *Sales Forecasting using SARIMAX for B2C* (pp. 1–5). https://doi.org/10.1109/icdsaai59313.2023.10452574

Nafouanti, B., Li, J., Nyakilla, E., Mwakipunda, C., & Mulashani, A. (2023). A novel hybrid random forest linear model approach for forecasting groundwater fluoride contamination. Environmental Science and Pollution Research, 30(17), 50661–50674. https://doi.org/10.1007/s11356-023-25886-w

Nagpal, A., & Gabrani, G. (2019). Python for Data Analytics, Scientific and Technical Applications. 2019 Amity International Conference on Artificial Intelligence (AICAI), 140-145. https://doi.org/10.1109/AICAI.2019.8701341

Naik, H., Yashwanth, S., & Jayapandian, N. (2022). Machine Learning based Food Sales Prediction using Random Forest Regression. IEEE Xplore. https://doi.org/10.1109/ICECA55336.2022.10009277

Navrátil, M., & Kolková, A. (2019). Decomposition and Forecasting Time Series in the Business Economy Using Prophet Forecasting Model. Central European Business review, 8, 26-39.

https://doi.org/10.18267/j.cebr.221.

Nguyen, N., Haider, M., Jisan, H., Raju, H., Imam, T., Khan, M., & Jafar, E. (2024). Product Demand Forecasting with Neural Networks and Macroeconomic Indicators: A Comparative Study among Product Categories. Journal of Business and Management Studies, 6(2), 170–175. https://doi.org/10.32996/jbms.2024.6.2.17

Nontapa, C., Kesamoon, C., Intrapaiboon, P., & Kaewhawong, N. (2020). A New Time Series Forecasting Using Decomposition Method with SARIMAX Model. Neural Information Processing, 743-751. https://doi.org/10.1007/978-3-030-63823-8_8

Noorunnahar, M., Chowdhury, H., & Mila, A. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. PLoS ONE, 18(3), e0283452. https://doi.org/10.1371/journal.pone.0283452

Nur, M., & Siregar, A. (2024). Exploring the Use of Cluster Analysis in Market Segmentation for Targeted Advertising. IAIC Transactions on Sustainable Digital Innovation (ITSDI). https://doi.org/10.34306/itsdi.v5i2.665

Okeleke, A., Ajiga, D., Folorunsho, O., & Ezeigweneme, C. (2024). Predictive analytics for market trends using AI: A study in consumer behavior. International Journal of Engineering Research Updates, 7(1), 036–049. https://doi.org/10.53430/ijeru.2024.7.1.0032

Ouyang, Y. Li, X., Zhou, W., Hong, W., Zheng, W., Qi, F., & Peng, L. (2024). Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology. PLoS ONE, 19(7), e0307478–e0307478. https://doi.org/10.1371/journal.pone.0307478

Qi, J., Du, J., Siniscalchi, M., Ma, X., & Lee, C. (2020). On mean absolute error for deep neural network based Vector-to-Vector regression. IEEE Signal Processing Letters, 27, 1485–1489. https://doi.org/10.1109/lsp.2020.3016837

Kluyver Thomas, Ragan-Kelley Benjamin, Pérez Fernando, Granger Brian, Bussonnier Matthias, Frederic Jonathan, Kelley Kyle, Hamrick Jessica, Grout Jason, Corlay Sylvain, Ivanov Paul, Avila Damián, Abdalla Safia, Willing Carol, & Jupyter Development Team. (2016). Jupyter Notebooks; a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press. https://doi.org/10.3233/978-1-61499-649-1-87

Ramos, R. F., Biscaia, R., Moro, S., & Kunkel, T. (2023). Understanding the importance of sport stadium visits to teams and cities through the eyes of online reviewers. Leisure Studies, 42(5), 693–708. https://doi.org/10.1080/02614367.2022.2131888

Rao, T., Sumanth, N., & Saiktishna, C. (2022). Historical Analysis and Time Series Forecasting of Stock Market using FB Prophet. 2022 6th International Conference on Intelligent

Computing and Control Systems (ICICCS), 1846-1851. https://doi.org/10.1109/ICICCS53718.2022.9788231.

Robeson, M., & Willmott, J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. PLoS ONE, 18(2), e0279774. https://doi.org/10.1371/journal.pone.0279774

Rosário, A., Moniz, B., & Cruz, R. (2021). Data science applied to marketing: A literature review. Journal of Information Science and Engineering, 37(5), 1067– 1081. Institute of Information Science. https://doi.org/10.6688/JISE.202109_37(5).0006

Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E., & Vidal, T. (2023). A Survey of Contextual Optimization Methods for Decision Making under Uncertainty. European Journal of Operational Research, 320(2), 271-289, https://doi.org/10.1016/j.ejor.2024.03.020

Sahin, O., & Kozat, S. (2019). Nonuniformly sampled data processing using LSTM networks. IEEE Transactions on Neural Networks and Learning Systems, 30(5), 1452– 1461. https://doi.org/10.1109/tnnls.2018.2869822

Savargiv, M., Masoumi, B., & Keyvanpour, R. (2021). A new random forest algorithm based on learning automata. Computational Intelligence and Neuroscience, *2021*, Article 5572781. https://doi.org/10.1155/2021/5572781

Seto, H., Oyama, A., Kitora, S., Toki, H., Yamamoto, R., Kotoku, J., Haga, A., Shinzawa, M., Yamakawa, M., Fukui, S., & Moriyama, T. (2022). Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-20149-z

Shah, I., Iftikhar, H., Ali, S., & Wang, D. (2019). Short-Term electricity demand forecasting using components estimation technique. Energies, 12(13), 2532. https://doi.org/10.3390/en12132532

Sharma, A., Patel, N., & Gupta, R. (2021). Enhancing Retail Sales Forecasting through LSTM Networks and ARIMA Models: A Comparative Analysis of AI Methodologies. https://doi.org/10.2478/amns-2025-0984

Sheridan, P., Wang, M., Liaw, A., Ma, J., & Gifford, M. (2016). Extreme gradient boosting as a method for quantitative Structure–Activity Relationships. Journal of Chemical Information and Modeling, 56(12), 2353–2360. https://doi.org/10.1021/acs.jcim.6b0059

Simmons, R. (2001). Book Review: Winners and Losers: The Business Strategy of Football. Journal of Sports Economics, 2, 379 - 381. https://doi.org/10.1177/152700250100200406.

Sousa, M., Tomé, M., & Moreira, J. (2022). Long-term forecasting of hourly retail customer flow on intermittent time series with multiple seasonality. Data Science and Management, 5(3), 137–148. https://doi.org/10.1016/j.dsm.2022.07.002

Stefenon, F., Seman, O., Mariani, C., & Coelho, S. (2023). Aggregating prophet and seasonal trend decomposition for time series forecasting of Italian electricity spot prices. Energies, 16(3), 1371. https://doi.org/10.3390/en16031371

Stroebel, T., Woratschek, H., & Durchholz, C. (2019). Clothes make the fan: The effect of team merchandise usage on team identification, fan satisfaction and team loyalty. Journal of Global Sport Management, 6(2), 185–202. https://doi.org/10.1080/24704067.2018.1531354

Taylor, J., & Letham, B. (2018). Forecasting at Scale. The American Statistician, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

UEFA (2023). The European Club finance and investment landscape. Retrieved from https://ecfil.uefa.com/2023

Tolulope, O., & Oguanobi, V. (2024). Data-driven strategies for business expansion: Utilizing predictive analytics for enhanced profitability and opportunity identification. International Journal of Frontiers in Engineering and Technology Research, 06(02), 071–081. https://doi.org/10.53294/ijfetr.2024.6.2.0035

Tremblay, A., & Newman, J. (2014). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. Psychophysiology, 52(1), 124–139. https://doi.org/10.1111/psyp.12299

Tseng, F., Yu, H., & Tzeng, G. (2002). Combining neural network model with seasonal time series ARIMA model. Technological Forecasting and Social Change, 69(1), 71–87. https://doi.org/10.1016/s0040-1625(00)00113-x

Van Haaren, J., & Van Den Broeck, G. (2011). Relational Learning for Football-Related Predictions. Latest Advances in Inductive Logic Programming, (pp. 237–244). World Scientific. https://doi.org/10.1142/9781783265091_0025

Vashishtha, K., Burman, V., Kumar, R., Sethuraman, S., Sekar, R., & Ramanan, S. (2020). Product age based demand forecast model for fashion retail. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2007.05278

Wang, W., & Lu, Y. (2018). Analysis of the mean Absolute Error (MAE) and the Root Mean square Error (RMSE) in assessing rounding model. IOP Conference Series Materials Science and Engineering, 324, 012049. https://doi.org/10.1088/1757-899x/324/1/012049

Wuttke, L. (2023). CRISP-DM Standard. Datasolut. https://datasolut.com/crisp- dm-standard/

Xu, Y., Jiang, S., Yan, J., Cheng, Q., Xiao, Y., Wang, C., Yan, J., & Wang, X. (2021). LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biology, 22. https://doi.org/10.1186/s13059-021-02492-y.

Yan, W., Yuan, Y., Yang, M., Zhang, P., & Peng, K. (2023). Detecting the risk of bullying victimization among adolescents: A large-scale machine learning approach. Computers in Human Behavior, 147, 107817. https://doi.org/10.1016/j.chb.2023.107817

ZeroZero (n.d.). ZeroZero - Porque todos os jogos começam assim. https://www.zerozero.pt/

Zou, H., & Yang, Y. (2003). Combining time series models for forecasting. International Journal of Forecasting, 20(1), 69–84. https://doi.org/10.1016/s0169-2070(03)00004-9

Žunić, E., Korjenić, K., Hodžić, K., & Donko, D. (2020). Application of Facebook's Prophet algorithm for successful sales forecasting based on real-world data. International Journal of Computer Science and Information Technology, 12(2), 23–36. https://doi.org/10.5121/ijcsit.2020.12203