Forecasting Analysis On Sport E-Commerce

Data Science Lab. Project

Vittorio Haardt - 853268
Luca Porcelli - 853189
Valentina Barbera - 856780
Luca Sinanaj - 844540

Contents

1 Introduction

Nowadays, it has become increasingly important for sales companies to go beyond traditional brick-and-mortar stores if they want to remain relevant in the market. Simply relying on in-store sales is no longer sufficient for the vast majority of companies, as the majority of sales now come from the online market. As a result, companies need to explore various approaches to tap into this growing segment. One option for companies is to rely on third-party providers like Amazon to handle their online sales. While this can be a convenient solution, many companies prefer to have their own personal e-commerce platforms. By doing so, they can avoid additional costs associated with using a third-party platform and have more control over their online presence. Having a personal e-commerce platform also allows companies to handle their online operations more easily and freely, adapting to their specific needs and preferences. However, e-commerce is not just a source of sales for companies; it is also a valuable data source. The data generated from e-commerce transactions provides crucial insights that can inform important business decisions. Sales companies understand the significance of data analysis in driving their strategies, and therefore, they are highly interested in individuals who can effectively analyze and interpret sales data.

In line with this, our project will focus on analyzing data from an e-commerce platform specializing in sport articles. By comparing various sectors within this e-commerce, we aim to extract important business information that can guide the company in making informed decisions. By conducting a comprehensive examination of the data, we can uncover patterns, identify potential growth opportunities, optimize pricing strategies, and refine marketing campaigns. The insights gained from this analysis will enable sales company in the sport articles sector to stay ahead of the competition, tailor their offerings to customer preferences, and ultimately drive sales growth.

The results obtained from the analysis we conduct will serve as a prototype of the information that the sales department of a company can leverage. Sales managers are particularly interested in gaining insights into future expectations and sales data to make informed decisions about implementing promotions. By examining historical data and identifying trends, they can anticipate customer demands and strategically plan promotional campaigns to maximize sales.

Understanding future expectations is crucial for sales managers as it allows them to align their strategies with market trends and customer preferences. By analyzing data from the e-commerce platform, we can identify patterns and forecast potential shifts in demand. This information empowers sales managers to make proactive decisions on product offerings, pricing strategies, and marketing campaigns, enabling them to stay ahead of the competition and cater to evolving customer needs.

Additionally, the analysis helps sales managers prepare for future sales by providing valuable insights into inventory management. By studying sales data, they can determine which products are popular, forecast demand fluctuations, and plan their restocking efforts accordingly. This allows them to optimize inventory levels, avoiding stockouts or excess inventory that can negatively impact profitability. The results of our analysis provide sales managers with quantitative information, such as sales trends, product performance, and customer behavior, enabling them to make data-driven decisions. For example, they can identify the best-selling product categories, evaluate the effectiveness of different marketing campaigns, and assess the impact of pricing strategies on sales performance. Armed with this knowledge, sales managers can allocate resources more effectively, identify areas for improvement, and develop targeted strategies to drive sales growth.

In our project, we have taken the dataset from a prominent sport articles e-commerce site, aiming to conduct a comprehensive analysis and provide valuable insights for the future. The dataset encompasses sales data from 2013 to 2022, capturing information on revenue generated from sales and the corresponding sectors. To facilitate our analysis, we focused on identifying the most performing sector and employed forecasting techniques to predict future trends.

Our approach involved examining the historical sales data for each sector, aiming to understand their individual trends and growth patterns. By comparing the future development of these sectors, we aimed to develop sales strategies that would capitalize on those projected to perform well. Simultaneously, we sought to gain insights into specific sectors, discerning whether new emerging sectors were mere fads or had the potential to consistently outperform the established sports market. The main objective of our analysis was to predict future sales for the most performing sectors. By doing so, we aimed to equip businesses with crucial insights that would inform their strategic decision-making for the future. These insights would enable companies to formulate effective sales strategies, allocate resources efficiently, and adapt their product offerings to align with projected market trends.

In the following section, we will delve into the precise methodology employed to analyze the dataset. We will outline the techniques utilized to identify the most performing sector, discuss the forecasting models employed, and elaborate on the specific steps taken to ensure accurate predictions.

2 Data and Data Exploration

The dataset we utilized for our analysis comprises data from a sport e-commerce platform. It contains over 26,000 observations, with each row representing the sales amount for a specific category on a given day. The dataset is organized into three columns, namely:

- **date**: This column represents the date and time of the sales transactions, spanning from January 1, 2013, to May 1, 2023.
- **sector**: The sector column categorizes the sales into various sectors. The sectors included in the dataset are as follows: Archery, Martial arts, Child, Baseball, Basketball, Vouchers, Football, Casual, Cycling, Dance, Fitness, Darts, Golf, Underwear, Sea, Swim, Padel, Skates, Fishing, Table Tennis, Rugby, Running, Skiing, Skateboard, Snowboarding, Soft Air, Diving, Tennis, Trakking, and Volley.
- **total**: The total column indicates the monetary value generated by each sale.

Prior to our analysis, the dataset had undergone preliminary cleaning. We primarily focused on removing rows that exhibited negative sales amounts or null values. These instances were likely attributed to typographical errors or missing data. By applying this cleaning procedure, we effectively eliminated approximately 100 rows from the dataset, ensuring the integrity and accuracy of the remaining observations. With the cleaned dataset, we were able to proceed with our analysis, confident in the quality and reliability of the data. The dataset's extensive timeframe and comprehensive sector categorization provided a robust foundation for our subsequent analysis and forecasting techniques.

To identify the top-performing sectors within the dataset, our approach involved studying the cumulative gains of each sector. We initially set a criterion of selecting only those sectors that had more than 100 sales days in the last available years. Consequently, we narrowed down the sectors to the following: Football, Casual, Cycling, Fitness, Padel, Fishing, Running, Skiing, Trekking, Tennis, Kids, and Snowboarding.

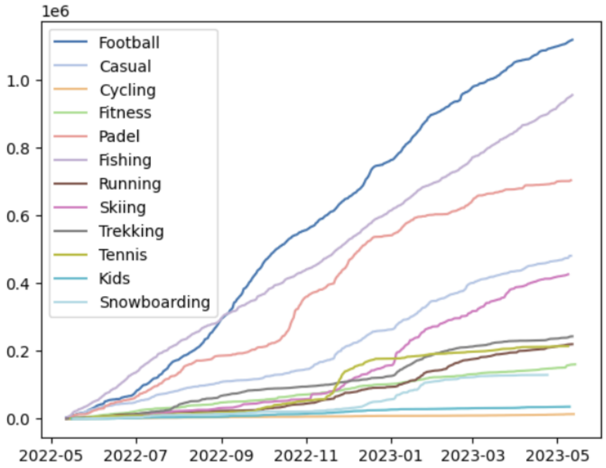

Figure 1: Cumulative gains.

To further refine our selection and determine the top five performing sectors, we plotted the cumulative gains for each sector. Figure 1 clearly illustrates the performance curves of these sectors, enabling us to visually compare their overall performance.

Upon analyzing the plot, it became evident that the sectors exhibiting consistently higher cumulative gains were *Football*, *Fishing*, *Casual*, *Padel*, and *Skiing*. These sectors consistently outperformed the others, displaying robust and upward trends in terms of cumulative gains over the analyzed period. Based on this observation, we identified these five sectors as the top performers for further analysis.

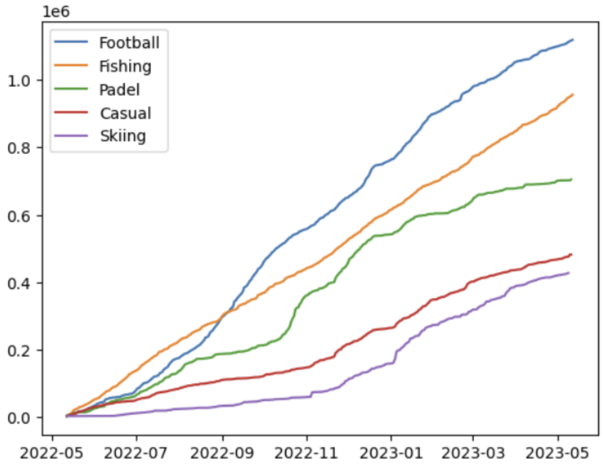

Figure 2: Cumulative gains of the 5 best performing sectors.

By focusing on these top-performing sectors, we aimed to gain deeper insights into their specific characteristics, trends, and potential growth opportunities. Analyzing their individual sales patterns, customer preferences, and market dynamics would enable us to develop targeted strategies and make informed business decisions.

After selecting the five top-performing sectors, we proceeded to plot their respective time series in order to gain initial insights simply by visually examining the data. However, due to the presence of noise in the data, we employed the technique of rolling mean to obtain a smoother representation of the time series.

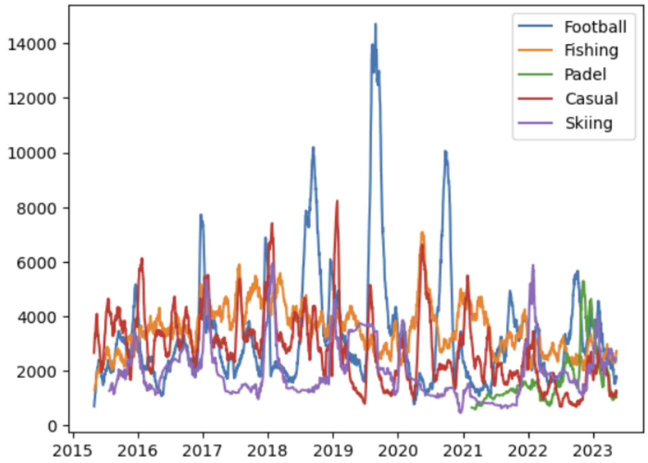

Figure 3: Smoothed time series of the five best performing sectors.

Figure 3 displays the plotted time series for each of the five sectors. While the plot may not appear immediately informative, it still provides valuable observations. For instance, we can observe that Padel sales only started to gain traction from 2021 onwards, indicating a strong upward trend in recent years. This suggests the emergence and growing popularity of this sport within the analyzed period. Moreover, it is evident that Football sales consistently exhibit the highest peaks among the five sectors. Interestingly, it appears that this sector was less affected by the 2019 pandemic compared to the others. While the sales peaks for the other sectors were relatively lower in the period from 2019 to 2020, Football sales remained robust.

In the subsequent sections, we will delve deeper into each individual sector to conduct a more comprehensive analysis of their respective features and trends. By analyzing each sector individually, we aim to extract valuable insights and make accurate predictions. This detailed examination will enable us to identify specific factors driving the sales performance of each sector and provide informed recommendations for future strategies.

3 Temporal Analysis

3.1 Introduction

The following analysis for each sector we will show how we conducted in a systematic manner, following a well-structured approach. Initially, the data was aggregated on a monthly basis to capture the sales patterns at that level. This allowed us to observe how sales fluctuated within each month over time, facilitating the identification of potential seasonality or trends. The analysis was then expanded to encompass sales per year, resulting in a chart that enabled easy comparison of monthly trends across different years. This chart provided valuable insights into the dynamics of sales within each month, allowing us to discern recurring patterns or shifts.

To gain a broader perspective of the overall sales growth and performance, we constructed a chart illustrating the annual cumulative sales. This visualization offered a clear representation of the total sales accumulated over each year, providing a sense of the overall sales trajectory and performance for each sector.

Another important aspect of the exploratory analysis was determining the historically highest revenue month for each specific category. This involved computing the cumulative sum of sales values for each individual month across multiple years. By doing so, we were able to identify the month that consistently recorded the highest revenue for a given category, offering valuable insights into seasonal peaks or specific periods of heightened sales activity.

The second part of the exploratory analysis focused on assessing the stationarity of the historical time series. Stationarity, characterized by consistent statistical properties over time, is a critical factor for selecting an appropriate forecasting model. To examine stationarity, we generated a chart that visualized the trend of the time series. By aggregating the data on a weekly basis, we could observe any prominent patterns or trends present, aiding in the identification of seasonality or other significant temporal patterns.

Furthermore, we paid attention to studying the autocorrelation among the values of the historical time series. Autocorrelation analysis helps uncover significant correlations between the current observation and lagged observations. This exploration provided insights into the temporal dependencies and relationships within the time series data, contributing to the selection of appropriate forecasting models and aiding in accurate predictions.

By conducting these comprehensive exploratory analyses, we were able to gain a deep understanding of the sales patterns, identify influential factors, and determine the most suitable forecasting models for each sector. This in-depth analysis paved the way for valuable insights, enabling us to make informed business decisions and develop effective sales strategies based on the historical trends and patterns exhibited by each sector.

After conducting the exploratory analysis, we proceeded to perform a forecasting analysis using three different models: ARIMA, SARIMA, and Prophet. The objective was to determine the most suitable model for each sector's time series data in order to generate accurate predictions.

The ARIMA (Autoregressive Integrated Moving Average) model, which incorporates autoregressive, moving average, and differencing components, was applied to each sector's data. Additionally, we employed the SARIMA (Seasonal ARIMA) model to account for any seasonal patterns present in the data. This model extends the capabilities of ARIMA by considering seasonal components. We also utilized the Prophet model, developed by Facebook's Core Data Science team. Prophet is a robust forecasting model that captures both trend and seasonality in the data. It incorporates additional features such as holiday effects, making it a versatile choice for analyzing various types of time series data. By comparing the performance of these models, we determined the model that provided the most accurate predictions for each sector. Once the best-performing model was identified, we proceeded to generate predictions for future sales based on the identified trends and patterns within the time series data.

To gain a deeper understanding of the potential range of future sales scenarios, we employed Monte Carlo Simulation. This technique allowed us to simulate multiple plausible paths or scenarios that the sales series could follow. By simulating various outcomes, we obtained a broader perspective on the potential future trajectories of sales.

The combination of the forecasting models and Monte Carlo Simulation enabled us to generate predictions for future sales and explore different scenarios. These insights proved invaluable for sales managers and decision-makers, as they provided valuable information on the potential future performance of each sector. With this knowledge, sales managers could make informed decisions regarding inventory management, marketing strategies, and resource allocation, ensuring they are well-prepared to meet future demand and capitalize on sales opportunities.

3.2 Football

3.2.1 Exploratory Analyses

Now, let's delve into the analysis of the Football sector, which stands out as the top-performing sector among all the others. Our initial focus was to observe the temporal time series of Football sales, utilizing moving average smoothing to better understand the trends during the considered period.

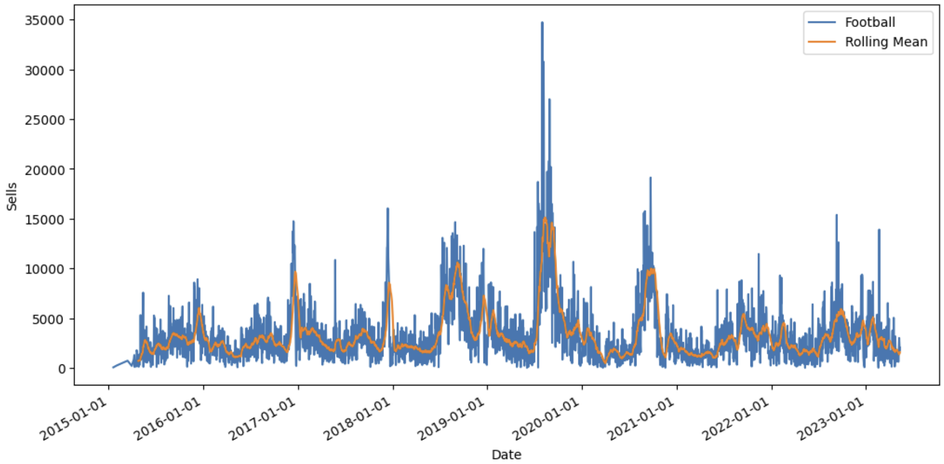

Figure 4: Football time series with the rolling mean smoothness.

In Figure 4, the temporal time series plot of Football sales reveals some noise in the data. To mitigate this noise and gain a clearer view, we decided to aggregate the data on a weekly basis. This aggregation method involves summing the sales data for each week, condensing the information while preserving the overall patterns and trends.

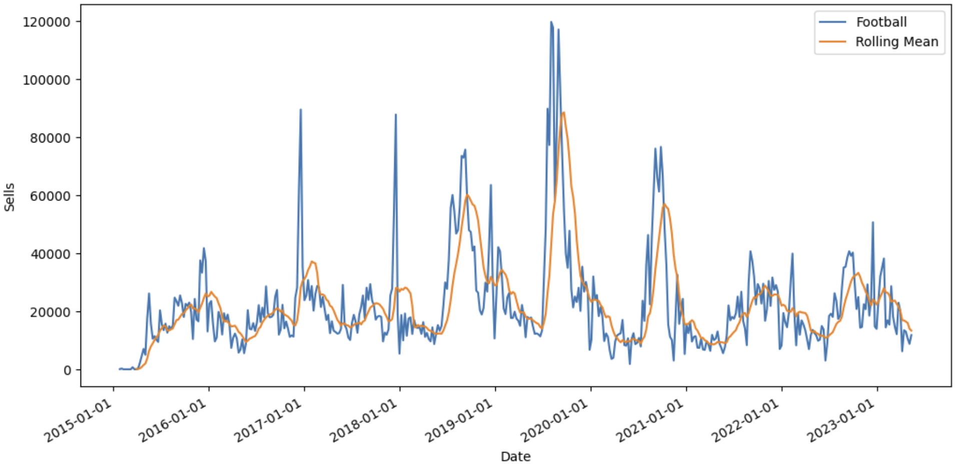

Figure 5: Football time series aggregated by week with the rolling mean smoothness.

Figure 5 displays the aggregated weekly trend for Football sales. Observing this chart, we notice that there is no significant upward or downward trend in the overall sales. The series appears relatively stable over time, with occasional peaks observed each year, particularly between 2019 and 2021. It is worth noting that the trend in the last two years appears to be less volatile compared to the preceding period.

By analyzing the aggregated weekly data, we gain a clearer understanding of the trend in Football sales. Although the series does not exhibit a pronounced upward or downward trajectory, the periodic peaks suggest potential seasonal patterns or external factors influencing sales. Furthermore, the relative stability observed in recent years may indicate a more consistent demand for Football-related products.

Following the initial visualization of the Football sector's time series, as previously mentioned, we proceeded to plot graphs representing the cumulative series for each year and each month.

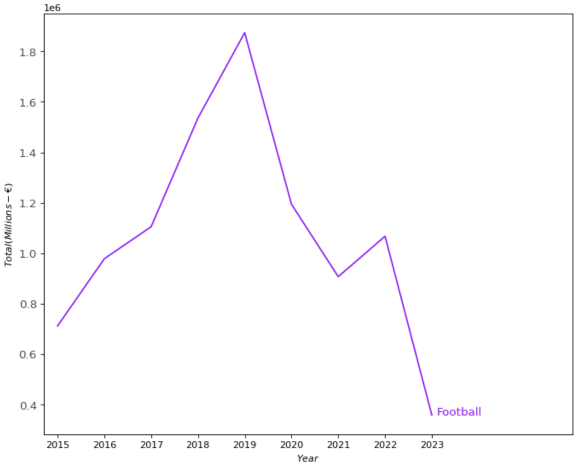

Figure 6: Cumulative gain among different years.

In Figure 6, we present the cumulative values of sales for each year. It is evident that sales consistently increased each year until 2019, reaching a peak. However, following this peak, there was a notable reversal in the trend, with sales declining by a magnitude similar to the previous increase. Notably, in 2022, we observe a renewed increment in sales. It is important to note that the cumulative value for 2023 should be interpreted with caution, as the data is only available for the months of January to May.

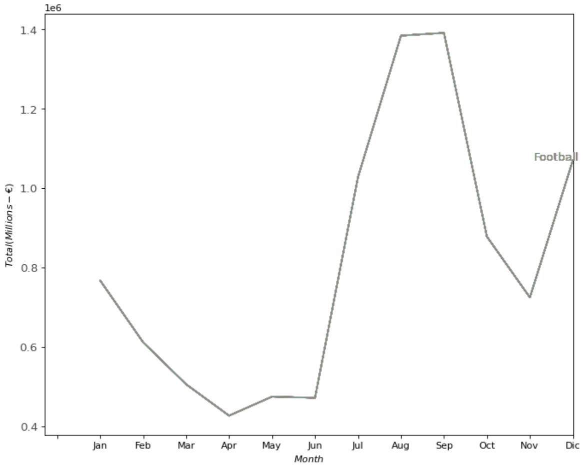

Figure 7: Cumulative gain for moth across different years.

Moving on to Figure 7, we depict the cumulative sales for each month, aggregating the sales figures for the same month across different years. This graph enables us to discern monthly trends. It is important to acknowledge that this type of analysis may introduce biases. For instance, if a particular month exhibits unusually high sales in a single year, it might skew the overall interpretation of that month's performance throughout the entire time series. To address this concern, we also provide cumulative figures for each month, segregated by year.

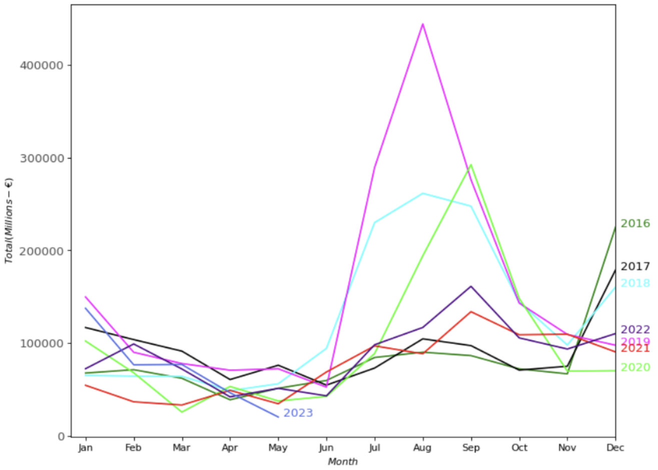

Figure 8: Cumulative gain for moth plotted for every year.

Examining Figure 8, we observe that the trend for the same month remains relatively stable across the entire time series, except for the period from July to September. During this period, particularly between 2018 and 2020, a significant surge in sales is evident, with 2019 standing out as a particularly exceptional year. Overall, the sales appear to be concentrated in the summer and winter months. Additionally, it is worth noting that while the sales in 2023 started off well, they have recently underperformed.

This analysis has provided valuable insights into the trend of Football sales over the years. We have observed a period of strong sales performance from 2018 to 2020, with notable exceptional sales during the summer months. However, in the most recent year, sales have experienced a slowdown. Understanding these trends and patterns is crucial for developing effective sales strategies, identifying potential influencing factors, and making informed decisions for future sales in the Football sector.

3.2.2 Forecasting

After completing the exploratory analysis, we transitioned to the core of our project: the forecasting analysis.

One of the key considerations in forecasting is determining whether the time series is stationary. Stationarity refers to the absence of trends, seasonal patterns, and other features commonly observed in real economic time series. A stationary time series exhibits consistent statistical properties over time, such as a constant mean and variance.

To assess the stationarity of the historical series under consideration, we employed two specific tests: the Augmented Dickey-Fuller (ADF) test and the KwiatkowSkiing-Phillips-Schmidt-Shin (KPSS) test.

The ADF test is a statistical procedure used to determine whether a time series exhibits a unit root, indicating nonstationarity. It compares the null hypothesis of a unit root's presence against the alternative hypothesis of stationarity. The ADF test is based on a regression model that includes the series' lags and their differences. The test evaluates the coefficient of the lagged variable in the regression model, and if the coefficient significantly differs from zero, there is evidence supporting the alternative hypothesis of stationarity. The ADF test follows the same testing procedure as the Dickey-Fuller test but is applied to the model:

$$y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \epsilon_t$$

where α is a constant, β is the coefficient on a time trend, p is the lag order of the autoregressive process, and γ represents the null hypothesis of $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. The unit root test is then conducted based on the computed test statistic $DF = \hat{\gamma}/SE(\hat{\gamma})$, which can be compared to the critical value for the Dickey-Fuller test. If the calculated test statistic is smaller (more negative) than the critical value, the null hypothesis of $\gamma = 0$ is rejected, indicating the absence of a unit root.

The KPSS test, on the other hand, is utilized to assess the stationarity of a time series. It compares the null hypothesis of stationarity against the alternative hypothesis of nonstationarity. The test computes a test statistic based on the cumulative sum of deviations from the series' mean. This statistic is then compared to critical values corresponding to the chosen significance level, such as 0.01 or 0.05. If the test statistic exceeds the corresponding critical value, the null hypothesis of stationarity is rejected, indicating that the series is nonstationary.

The KPSS test can be used as a complement to the ADF test, which focuses on the alternative hypothesis of a unit root's presence. Both tests provide valuable information for evaluating the stationarity of a time series. In addition to these tests, we decided to complement our analysis with descriptive plots to visually inspect the obtained results and determine the stationarity of the series.

We examined various plots to identify any seasonal patterns. However, based on our observations, we did not find any noticeable seasonal patterns. One particular plot we relied on was the partial autocorrelation graph, as depicted in the figure below.

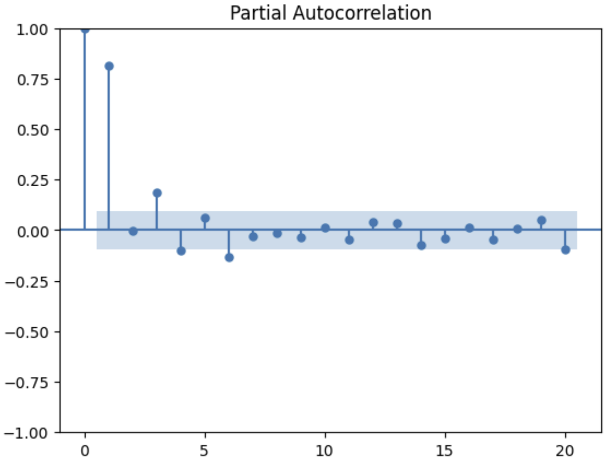

Figure 9: Partial autocorrelation plot

The key characteristic of this type of graph is the presence of a blue area, which represents the 95% confidence interval and serves as a significance threshold. Anything falling within the blue area is statistically close to zero, while values outside the blue area are considered statistically non-zero. On the X-axis, we have different types of lags, with a lag of 0 indicating the correlation of the time series with itself, which is always 1. By examining the graph, we can infer that the majority of autocorrelations are close to zero, suggesting that the time series appears to be stationary. Furthermore, both tests we conducted, namely the ADF Test and the KPSS Test, confirmed the stationarity of this historical series. These tests provided additional evidence supporting the absence of unit roots and the presence of stationarity in the soccer sector data.

Upon examining the time series, we found that it exhibited clear stationarity. As a result, we proceeded to utilize models specifically designed for stationary data. For model validation purposes, we set aside a 52-week (1-year) period to evaluate the performance of the selected models. Utilizing the weekly aggregated data, we fit three different models: ARIMA, SARIMA, and Prophet.

To identify the optimal parameters for each model, we employed a step function that evaluated the Akaike Information Criterion (AIC). The AIC provides a measure of the relative quality of a statistical model, taking into account both goodness of fit and model complexity. By selecting the model with the lowest AIC value, we obtained the most suitable parameters for each model.

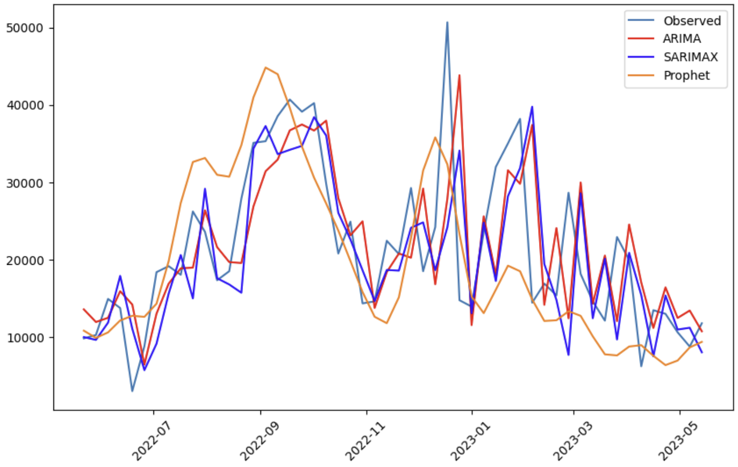

Figure 10: ARIMA, SARIMA and Prophet model fitting the validation period.

After fitting the models, we evaluated their performance on the test data using the Mean Squared Error (MSE) as the evaluation metric. The MSE quantifies the average squared difference between the predicted values and the actual values. While the MSE may appear relatively high in absolute terms, it is important to consider the magnitude of the sales data in this context. The results of the evaluation are presented in Figure 10 and the accompanying table. The MSE values for each model were as follows:

Model	MSE
ARIMA	8799.70
SARIMA	8531.41
Prophet	8513.37

Despite the relatively high MSE values, it is crucial to note that the magnitude of the sales data impacts these values. Upon visual inspection, we observed that all three models adequately captured the trend and patterns of the validation period. However, based on the lowest MSE, we selected the Prophet model as the most suitable for forecasting the Football sector sales.

After selecting the Prophet model as the most suitable for forecasting the Football sector sales, we proceeded to compute the 95% confidence intervals to evaluate the potential range of future values.

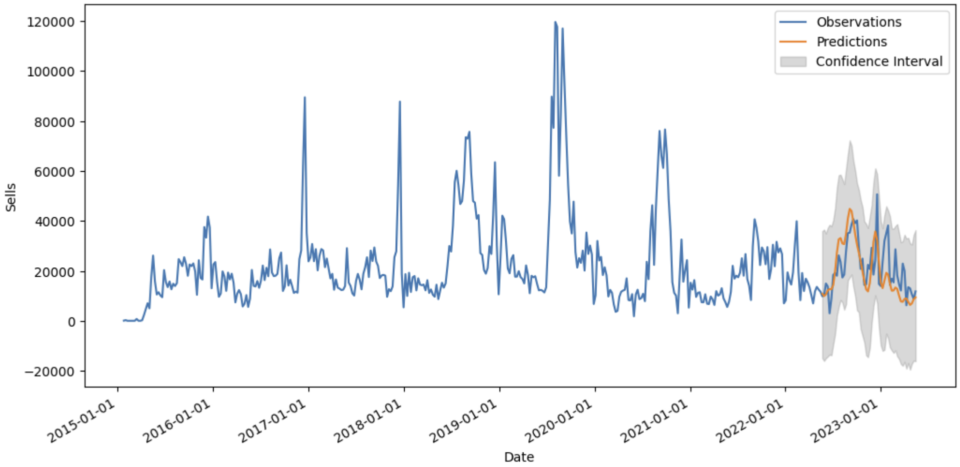

Figure 11: The best model (Prophet) predictions and their relative confidence interval at 95%.

In Figure 11, the forecasted values on the validation are represented by the orange line, while the shaded area represents the confidence intervals. It is important to note that the intervals appear relatively wide, but this is largely influenced by the high magnitude of the sales data.

Despite the broad intervals, the Prophet model demonstrates satisfactory performance and provides valuable insights for future projections. The model's ability to capture the underlying trends and patterns of the data instills confidence in its applicability for forecasting future sales in the Football sector.

3.2.3 Prediction

To enhance the robustness of our analysis and provide more comprehensive insights, we employed Monte Carlo simulations to generate 1000 different paths for the Football sector sales prediction. Each path was created by introducing random noise into the forecasting model. This approach allows us to obtain a range of potential outcomes rather than a single deterministic forecast. By simulating various paths, we can account for the inherent uncertainty and fluctuations in the sales data. While the individual paths may differ in direction and magnitude, collectively they provide a more comprehensive understanding of the potential future trajectories for the sector.

Given that the Football sector sales have remained relatively stable over the past 10 years without significant increases or decreases, the simulated paths may exhibit diverse directions. This reflects the inherent unpredictability and variability in the sales patterns.

To gain further insights from the Monte Carlo simulations, we analyzed the distribution of daily sales values and the cumulative distribution for the year 2023. These visualizations help us understand the range and variability of the simulated paths.

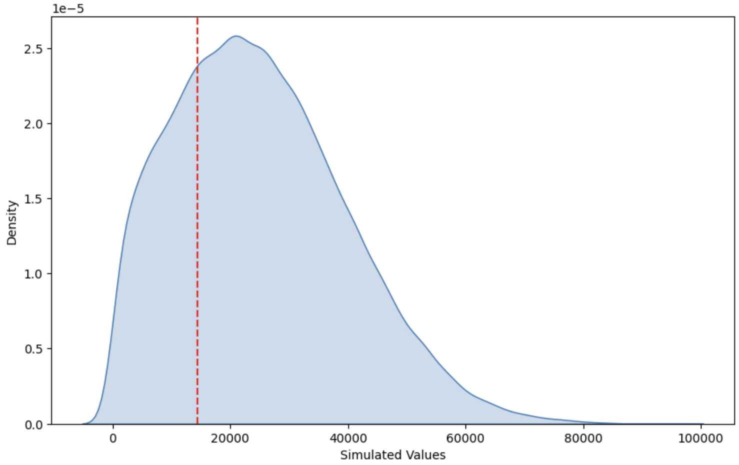

Figure 12: Distribution of daily sales for all the 1000 path.

Figure 12 represents the distribution of daily sales values. It provides an overview of the predicted values and their realism. The graph shows that the average daily sales values are below 20,000, and the distribution is asymmetric due to the presence of many days with zero sales. The range of values spans from 0 to a maximum of 80,000.

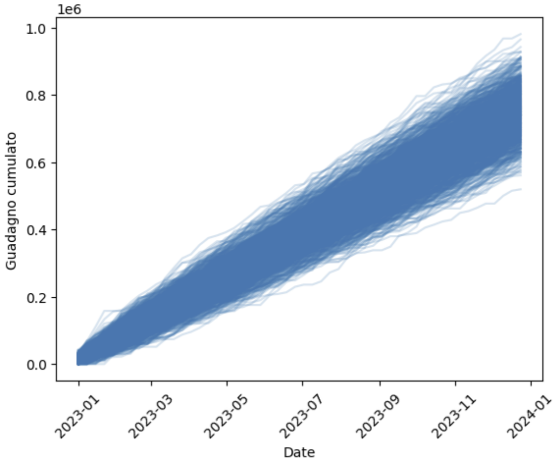

Figure 13: Cumulative curve for the 1000 simulated path.

Moving on to Figure 13, we observe the cumulative distribution for the year 2023, generated from the 1000 different simulated paths. The graph illustrates the range of possible outcomes for the total gain from Football sector sales in that year. The majority of the paths indicate cumulative gains between 650,000 and 850,000. This suggests that the cumulative sales for 2023 are expected to be lower than the previous year, further reverse the trend reversal observed in Figure 6.

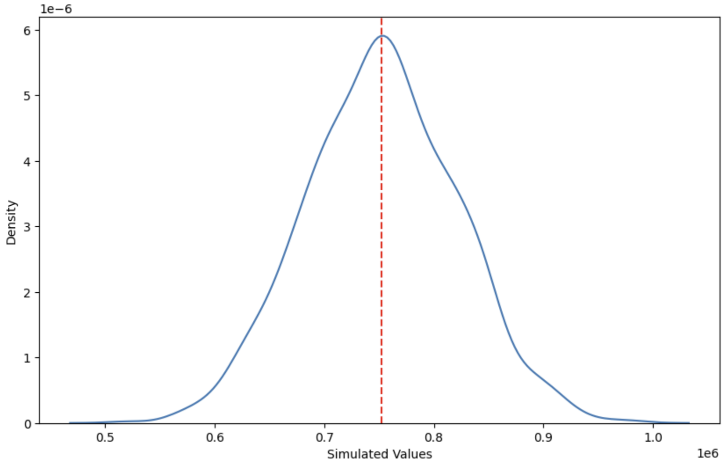

Figure 14: Distribution of the cumulative value at the end of the future year for 1000 simulated path.

In Figure 13, we present the density distribution of values across the different paths. This visualization provides a more detailed view of the results. We can observe that the values are concentrated around the mean value of 750,000. As mentioned earlier, the average annual decrease in Football sector sales is expected to be around 200,000, which is reflected in the density distribution.

By considering the range of possible scenarios and understanding the distribution of values, businesses can make more informed decisions and develop strategies to adapt to potential changes in the market.

3.3 Fishing

3.3.1 Exploratory Analyses

The next sector we analyzed is the Fishing sector, which also performed well in terms of sales. As shown in Figure 1, it even topped the charts in the early months. Similarly to our previous approach, we started by examining the temporal time series of sales using moving average smoothing.

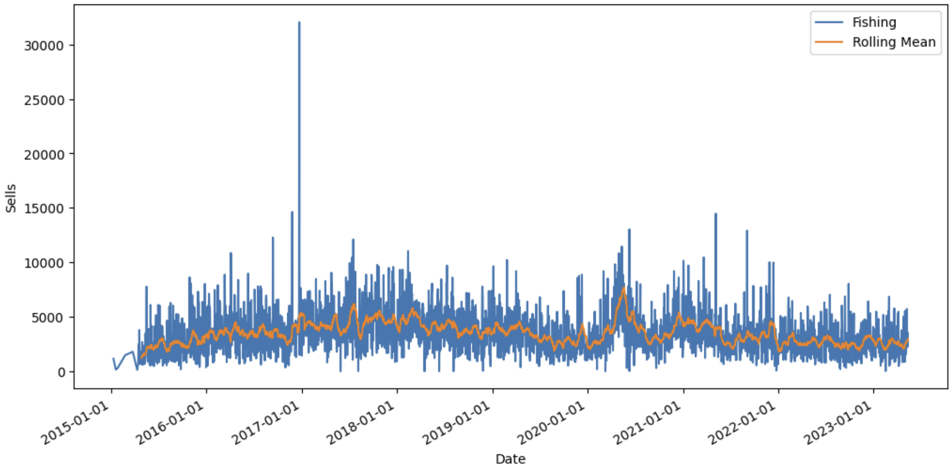

Figure 15: Fishing time series with the rolling mean smoothness.

Figure 15 provides a visual representation of the trend in Fishing sales. As with the other sectors, we decided to aggregate the data on a weekly basis to obtain a clearer view.

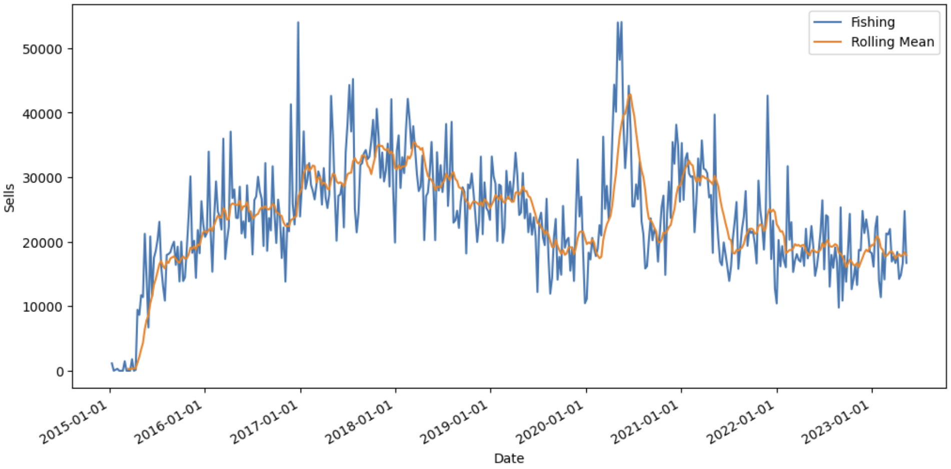

Figure 16: Fishing time series aggregated by week with the rolling mean smoothness.

Figure 16 presents the aggregated weekly trend for Fishing sales. It reveals that, similar to the previous sector, there is no significant upward or downward trend in overall sales. The series appears to be relatively stable over time, with occasional peaks and a minor decline at the beginning of 2020, followed by a strong recovery. By analyzing the aggregated weekly data, we gain a better understanding of the trend in Fishing sales. The series appears to be stable, suggesting the absence of seasonality or significant external influences on sales in this sector.

Moving on to the cumulative series, we plotted graphs to visualize the sales trends for each year and each month.

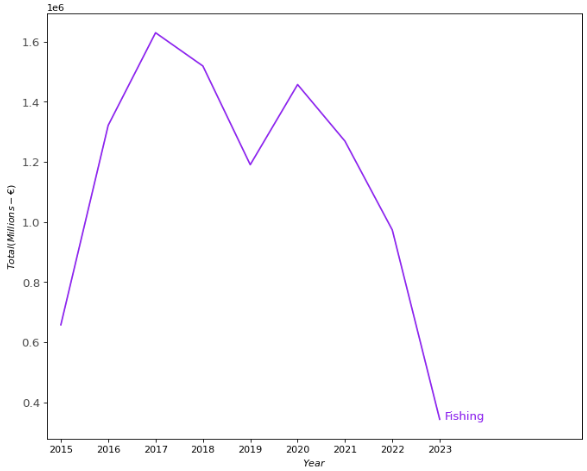

Figure 17: Cumulative gain among different years.

Figure 17 displays the cumulative values of sales for each year. It is evident that sales consistently increased each year until 2017, reaching a peak. However, a subsequent reversal in the trend occurred, with sales declining by a magnitude similar to the previous increase. Notably, in 2020, there was a renewed increment in sales, followed by a significant decrease, resulting in the worst performance of the series in 2022. We should interpret the cumulative value for 2023 cautiously, as the data is only available for the months of January to May.

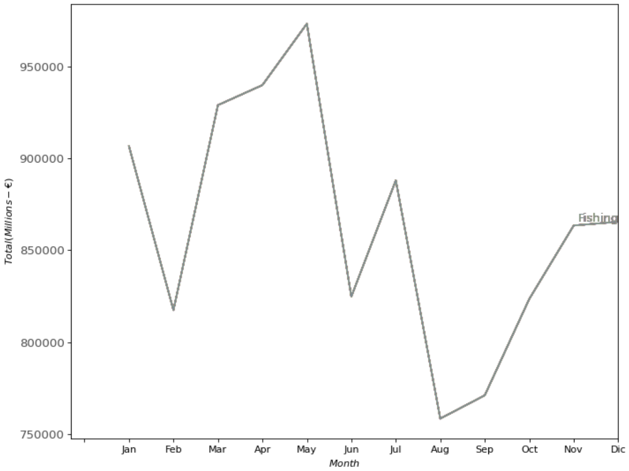

Figure 18: Cumulative gain for moth across different years.

Figure 18 focuses on the cumulative sales for each month, aggregating the sales figures for the same month across different years. It provides insights into the performance of sales throughout the year.

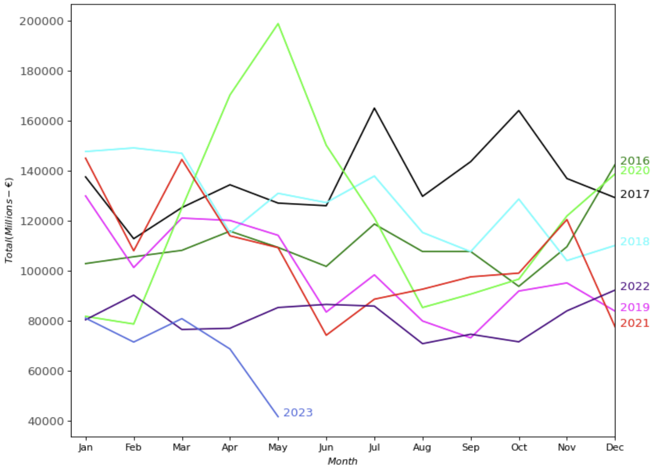

Figure 19: Cumulative gain for moth plotted for every year.

Figure 19 highlights the yearly trends, showing that each year has its unique sales pattern. It is important to note that the representation of May as a particularly profitable month may be influenced by the specific sales pattern observed in 2019, but examining other years reveals that it is not consistently the case. This plot also illustrates how the Fishing sector has experienced a decline in sales over the years.

3.3.2 Forecasting

In our analysis of the stationarity of the Fishing sector, the descriptive graphs did not reveal any clear seasonal patterns. Specifically, examining the Partial Autocorrelation Plot, we observed that autocorrelation remained significant from lag 0 to lag 5, but the majority of autocorrelations were found to be zero.

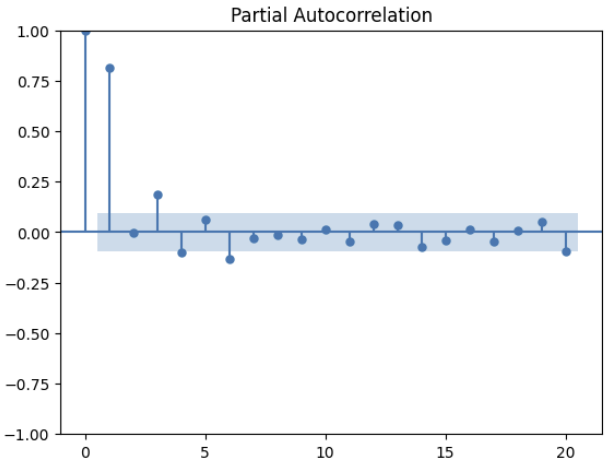

Figure 20: Partial autocorrelation plot

In order to better understand the stationarity of this series we can analyze the result of the tests. For the ADF test the stationarity hypothesis is confirmed but not for the KPSS test, which instead considers the series to be non-stationary. Because of these unclear results, we decided to further deepen the analysis by decomposing the time series, with the “decompose” function to go for seasonal patterns not easily recognized. After reapplying the tests to the decomposed series, the nonstationarity was confirmed.

Next we decided to apply the three models seen previously for the soccer sector, expecting a better performance for SARIMA or Prophet considering the nonstationarity of the series. In the following plot we can see SARIMA,ARIMA and Prophet model fitting the validation period.

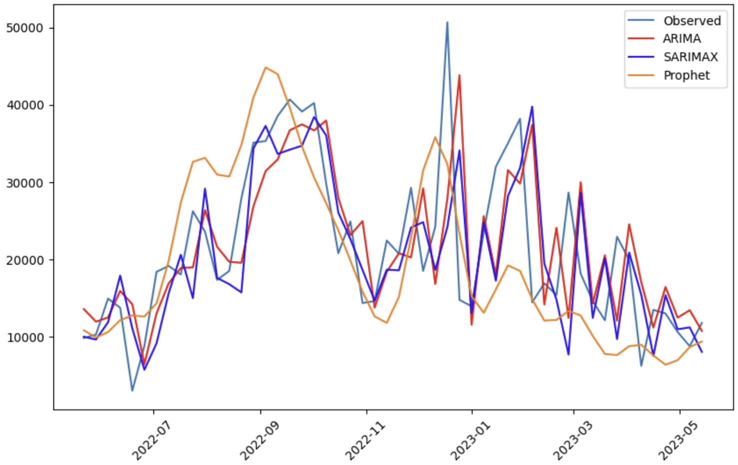

Figure 21: ARIMA, SARIMA and Prophet model fitting the validation period.

We can see that the red line of the Arima model is not visible because the performance is very similar to the performance of the Sarimax model.

Subsequently, we applied the three models discussed previously for the soccer sector, expecting a more favorable performance from either SARIMA or Prophet given the nonstationarity of the series. The model that exhibited superior performance was Prophet, with a mean squared error (MSE) of 3724, significantly lower than SARIMA and ARIMA. The MSE values for each model are presented in the following table:

Model	MSE
ARIMA	4504.24
SARIMA	4504.25
Prophet	3724.12

Having selected the Prophet model as the most suitable for forecasting Fishing sector sales, we proceeded to compute the 95% confidence intervals to assess the potential range of future values.

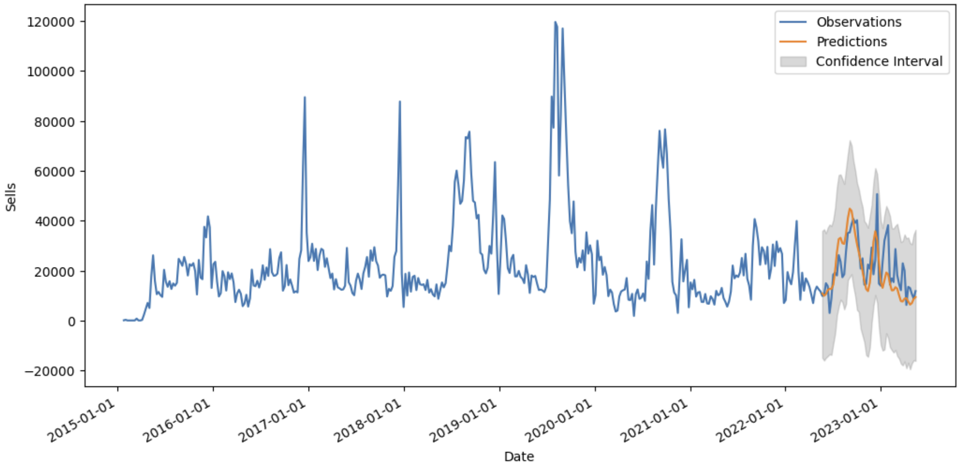

Figure 22: The best model (Prophet) predictions and their relative confidence interval at 95%.

It is important to note that the intervals appear relatively wide, but this is largely influenced by the high magnitude of the sales data. Despite the broader intervals, the Prophet model demonstrates satisfactory performance and provides valuable insights for future projections.

3.3.3 Prediction

As before, we used Monte Carlo simulations to generate 1000 different sales predictions in the Fishing sector. Each prediction was created by introducing random noise into the forecasting model. Since sales in the Fishing sector have remained relatively stable over the past 10 years without significant increases or decreases, the simulated paths can exhibit various directions. This reflects the unpredictability and intrinsic variability in sales models.

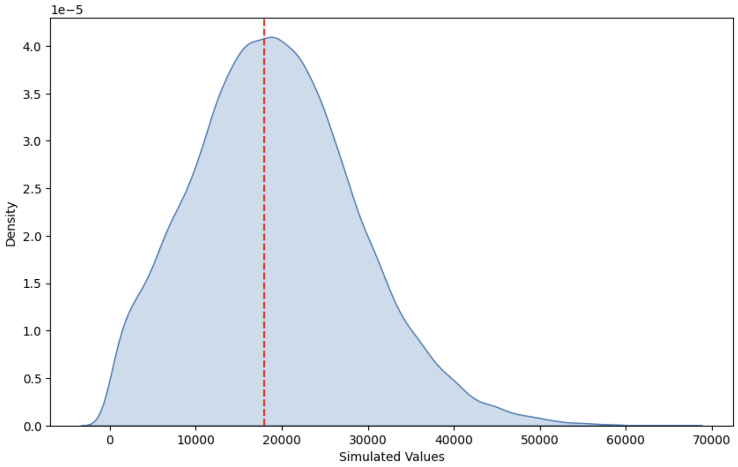

Figure 23: Distribution of daily sales for all the 1000 path.

Figure 23 represents the distribution of daily sales values in the Fishing sector. It can be observed that the average daily sales values remain below 20,000, while the distribution is characterized by a slight asymmetry, likely due to days with zero sales. The range of values spans from 0 to a maximum of 60,000.

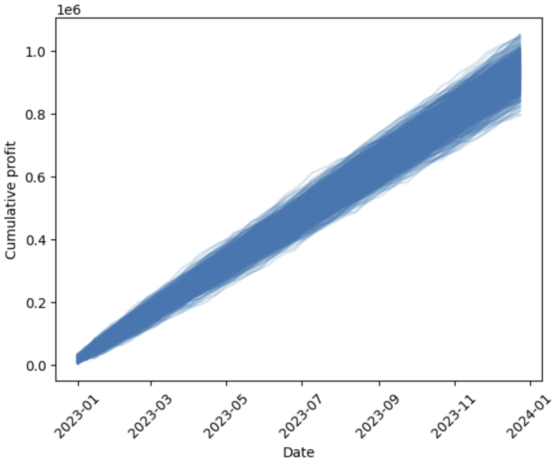

Figure 24: Cumulative curve for the 1000 simulated path.

In Figure 24, we present the cumulative distribution for the year 2023 in the Fishing sector, based on 1000 different simulations. The chart highlights the variety of possible outcomes for the total sales revenue in the sector during the year. In this case, there is a substantial similarity among the results, as the range is not wide and the cumulative revenue exceeds 800,000. This aspect supports what was mentioned in Figure 6, namely that a decrease in sales is projected for the upcoming year.

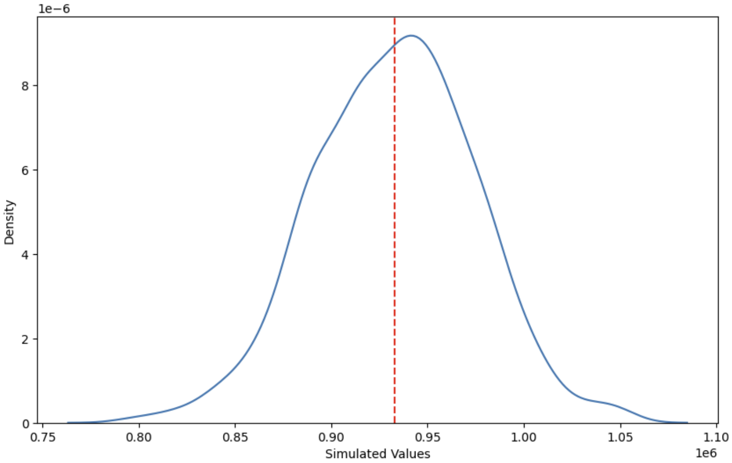

Figure 25: Distribution of the cumulative value at the end of the future year for 1000 simulated path.

Figure 25 presents the density distribution of values across different paths in the Fishing sector. It can be observed that the values are concentrated around the mean value of nearly 950,000. As mentioned earlier, an average annual decrease in Fishing sector sales of approximately 50,000 is expected, which is reflected in the density distribution.

3.4 Padel

3.4.1 Exploratory Analyses

Padel represents another sector that has experienced significant sales growth in the past year. The historical series for Padel sales is available from 2020, reflecting the sport's recent surge in popularity. To analyze the sales trends, we applied moving average smoothing to the temporal time series.

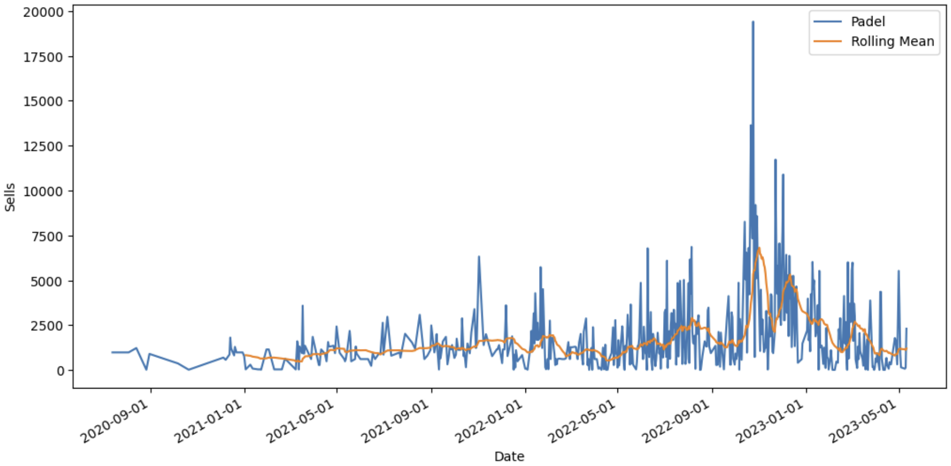

Figure 26: Padel time series with the rolling mean smoothness.

Here is the graph with weekly aggregation to reduce noise.

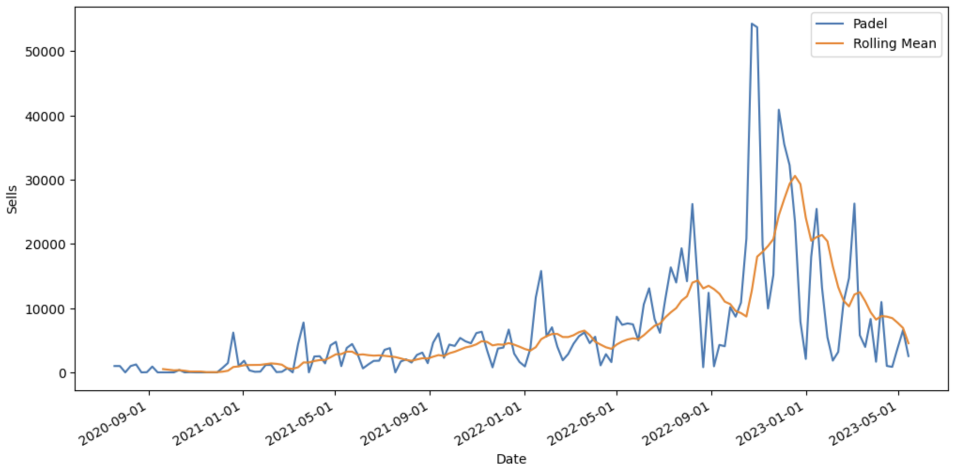

Figure 27: Padel time series aggregated by week with the rolling mean smoothness.

Figure 27 depicts the weekly aggregated trend for Padel sales, allowing us to reduce noise and identify underlying patterns. From this representation, we do not observe any specific trends in the sales data. However, there is a notable increase in sales starting from 2022 and continuing into 2023, reaching a peak. This suggests a strong and growing demand for Padel-related products in the upcoming years. Since the data is aggregated on a weekly basis, it is not possible to ascertain the presence of seasonality. To gain more insights into the historical series, we proceed with aggregating the data by month.

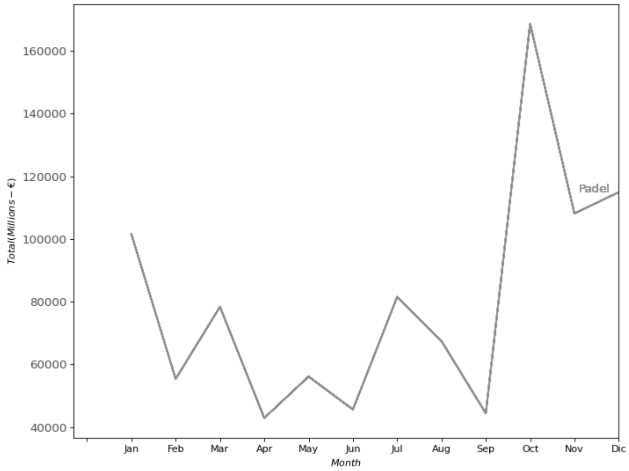

Figure 28: Cumulative gain for moth across different years.

Figure 28 presents the aggregated monthly trend for Padel sales. It reveals a noticeable increase in demand during the summer period, particularly from June onwards. However, it is important to note that the representation may be influenced by the significant sales surge in October 2022, which can be further verified in the subsequent graph displaying aggregated monthly sales segregated by year. As more years pass and additional data becomes available, we will be able to conduct more precise and accurate analyses to determine whether Padel is a passing trend or a sector worth investing in.

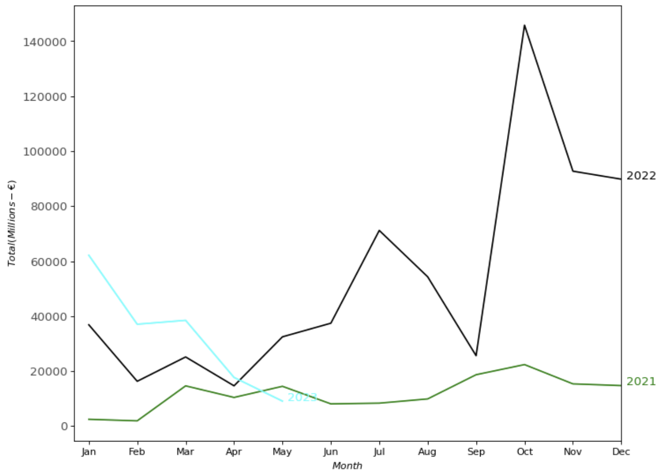

Figure 29: Cumulative gain for moth plotted for every year.

3.4.2 Forecasting

The Padel sector emerged around 2021, and from an initial descriptive analysis, we observed a sales peak between 2022 and 2023. At first glance, no clear seasonal patterns were evident. However, considering that this sector is relatively new, we conducted a more comprehensive analysis to identify potential non-stationarity.

Examining the Partial Autocorrelation plot below, we noticed that only a few autocorrelations were statistically significant, while the majority were close to zero. Thus, we can preliminarily consider this time series to be stationary. However, to gain a better understanding, we need to examine the results of the tests.

Again we applied three models ARIMA,SARIMA and Prophet and in the following plot we can see the performance on the validation period for each of them.

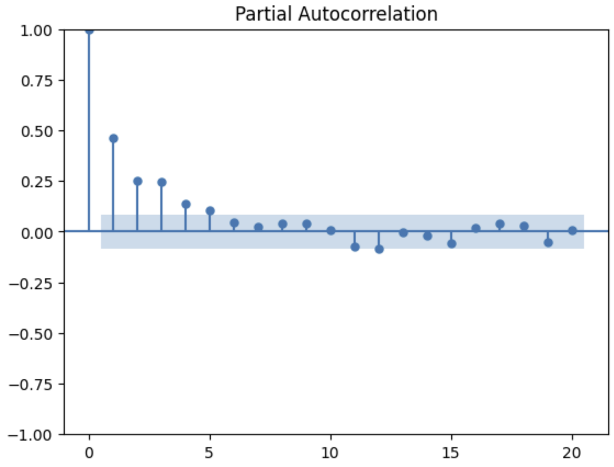

Figure 30: Partial autocorrelation plot

Also in this case the performance of ARIMA and Sarimax model are very similar so is not possible to distinguish the two lines.

Applying the tests to the non-decomposed historical series, the KPSS test indicated non-stationarity, similar to our findings for the Fishing sector. Consequently, we decomposed the historical series and re-applied the tests to the decomposed series, which confirmed its stationarity.

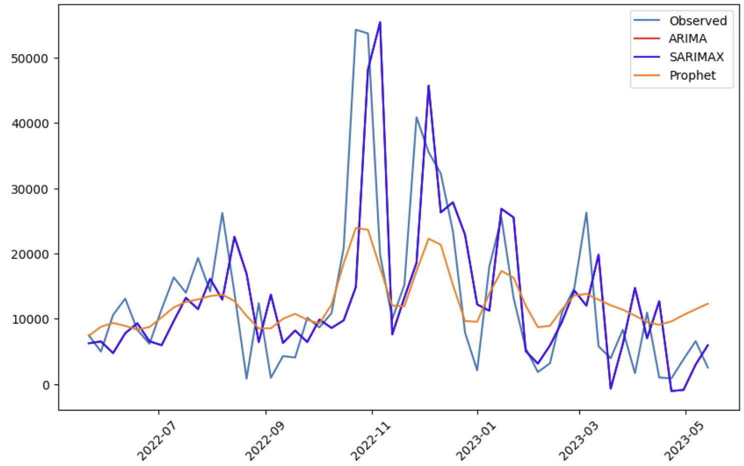

Figure 31: ARIMA, SARIMA and Prophet model fitting the validation period.

Next, we applied three models: ARIMA, SARIMA, and Prophet. Among these models, Prophet yielded the best performance with a mean squared error (MSE) of 9014. Notably, the performance difference between Prophet and ARIMA/SARIMA models was substantial. The following table displays the MSE values for each model:

Model	MSE
ARIMA	10688.05
SARIMA	10688.06
Prophet	9014.66

Having selected the Prophet model as the most suitable for forecasting Padel sector sales, we computed the 95% confidence intervals to assess the potential range of future values.

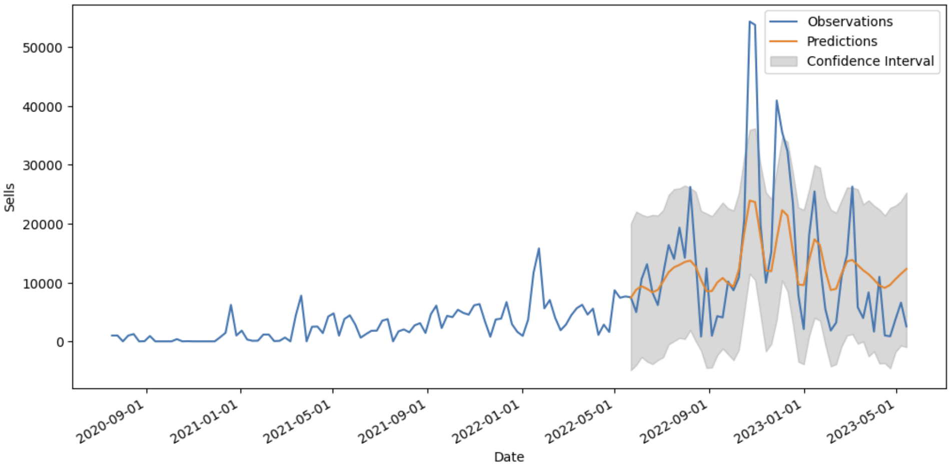

Figure 32: The best model (Prophet) predictions and their relative confidence interval at 95%.

It is important to note that the intervals appear relatively wide, primarily due to the high magnitude of the sales data. Despite these broader intervals, the Prophet model demonstrated satisfactory performance and provided valuable insights for future projections.

3.4.3 Prediction

Compared to the other sectors analyzed, the Padel sector has experienced a phase of strong growth in recent years. However, more recently, sales have shown a decline, returning to lower levels. Despite this decrease, the Padel sector has maintained a certain level of stability. Therefore, for the Padel sector as well, Monte Carlo simulations can generate a variety of paths, reflecting the unpredictability and intrinsic variability in sales models.

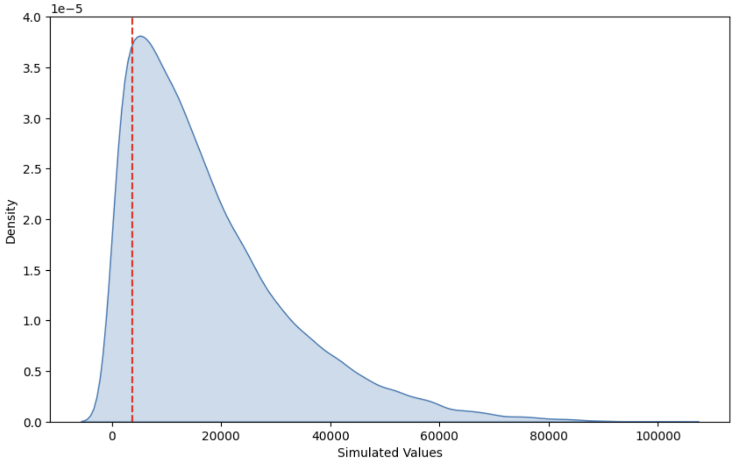

Figure 33: Distribution of daily sales for all the 1000 path.

Figure 33 displays the distribution of daily sales values in the Padel sector. This chart provides an overview of the predicted values and their realistic distribution. It can be noted that the average daily sales values are below 10,000. The distribution's asymmetry is quite evident due to numerous days with zero sales. The range of values varies from 0 to a maximum of 100,000.

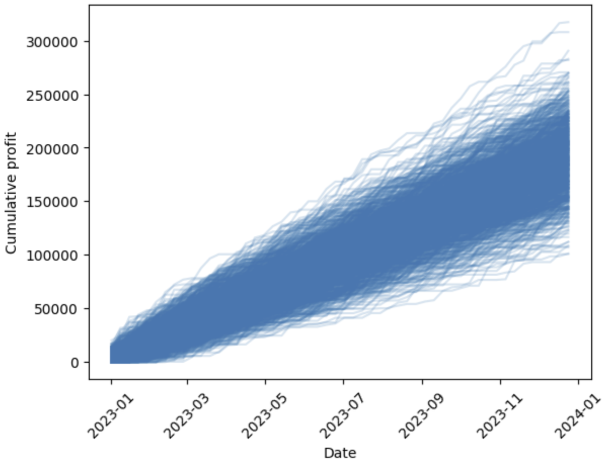

Figure 34: Cumulative curve for the 1000 simulated path.

In Figure 34, the cumulative distribution for the year 2023 in the Padel sector is presented, based on 1000 different simulations. As mentioned earlier, most trajectories indicate cumulative earnings, but unlike the Fishing sector, the Padel sector exhibits a wide range of possible solutions ranging from less than 100,000 to over 300,000. In this case, a significant decrease from the previous year is also observed, but it may be influenced by exceptional growth in the Padel sector in the previous year, as highlighted in Figure 8. However, unlike the other results, it may be more of a stabilization rather than a continuing downward trend.

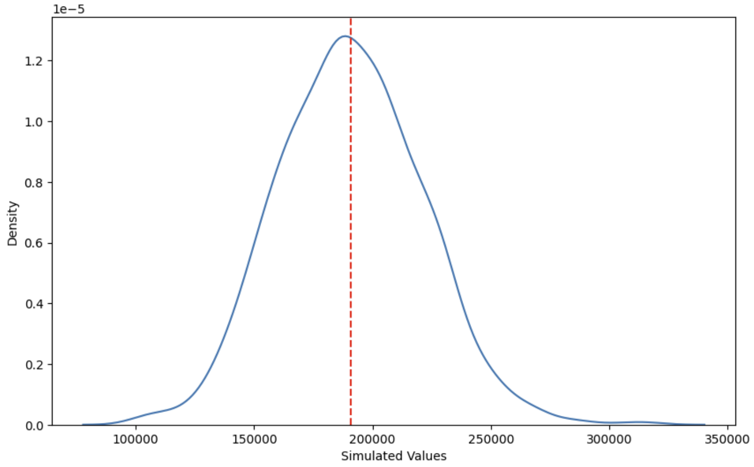

Figure 35: Distribution of the cumulative value at the end of the future year for 1000 simulated path.

As observed in Figure 35, the density distribution of values across different paths in the Padel sector is represented. It can be noted that the values are concentrated around the mean value of approximately 200,000. As previously mentioned, a significant average annual decrease in Padel sector sales of about 400,000 is expected. This decline can be attributed to the remarkable growth the sport experienced last year.

3.5 Casual

3.5.1 Exploratory Analyses

Now let's delve into the analysis of the fourth sector: Casual. Similar to our approach with the previous sectors, let's examine the time series using moving average smoothing.

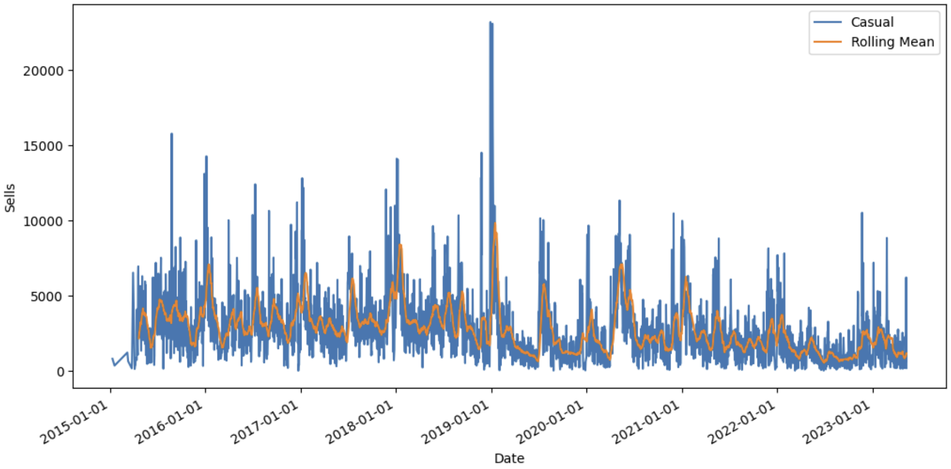

Figure 36: Casual time series with the rolling mean smoothness.

To further explore the time series of the Casual sector, we aggregated the data on a weekly basis to reduce noise.

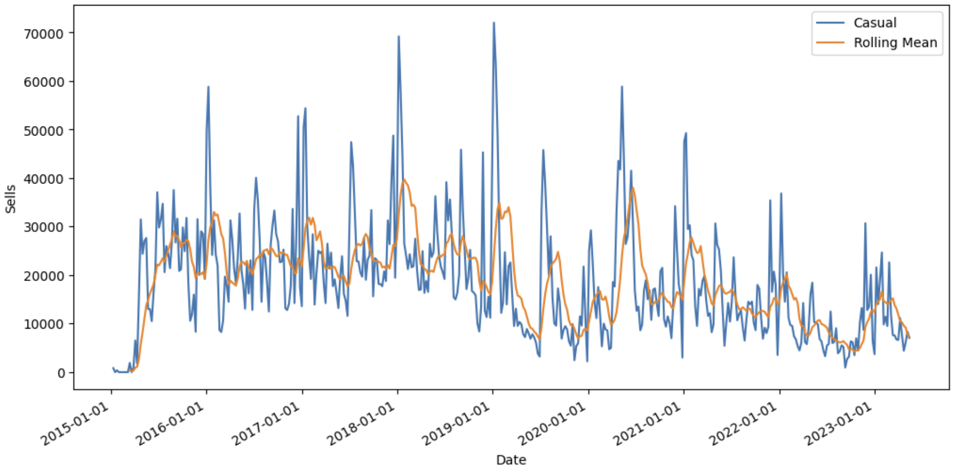

Figure 37: Casual time series aggregated by week with the rolling mean smoothness.

Figure 37 illustrates the aggregated weekly trend for Casual sales. We observe sporadic peaks without any discernible overall trend. However, it is important to note that the peak sales occur consistently in each year, especially in 2019. This indicates a seasonal component in the sales pattern of the Casual sector.

To gain more insights into the Casual sector's sales patterns, we proceeded to plot the cumulative series for each year and month.

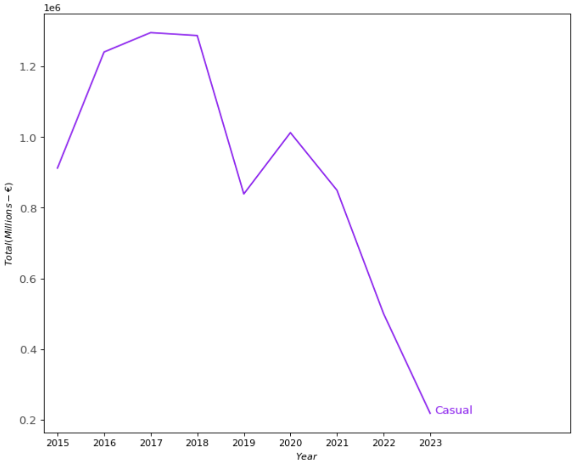

Figure 38: Cumulative gain among different years.

Figure 38 provides the cumulative sales values for each year, highlighting a significant decline in sales between 2021 and 2022, following a positive trend that began in 2019.

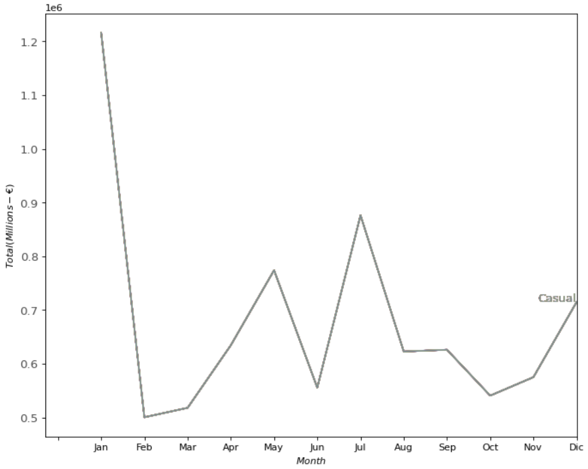

Figure 39: Cumulative gain for moth across different years.

Figure 39 presents the aggregated sales per month for the Casual sector, divided by year. As observed in the previous sectors, there are sales peaks in May and July. However, it is crucial to note that this representation can sometimes be misleading. To gain a clearer understanding, we examine the aggregated sales per month for each year.

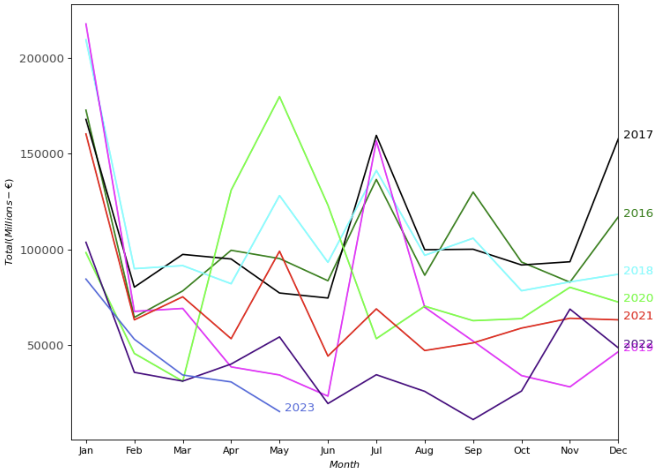

Figure 40: Cumulative gain for moth plotted for every year.

Figure 40 allows us to observe that almost every year experiences two sales peaks, occurring in May and July as previously mentioned. Additionally, there is a noticeable decline in sales during February. This analysis, combined with the previous findings, indicates the presence of a seasonal component in the time series of the Casual sector. This valuable information can guide the development of new sales strategies and facilitate a comprehensive evaluation of past strategies.

3.5.2 Forecasting

For the Casual sector from the descriptive graphs we do not notice at first glance the presence of seasonal patterns. From the following Partial Autocorrelation plot we can see that from lag 0 to lag 9 some the autocorrelations value are significant so we can hypothesis the presence of a non- stationary series , but in order to better understand this we need to see the tests results.

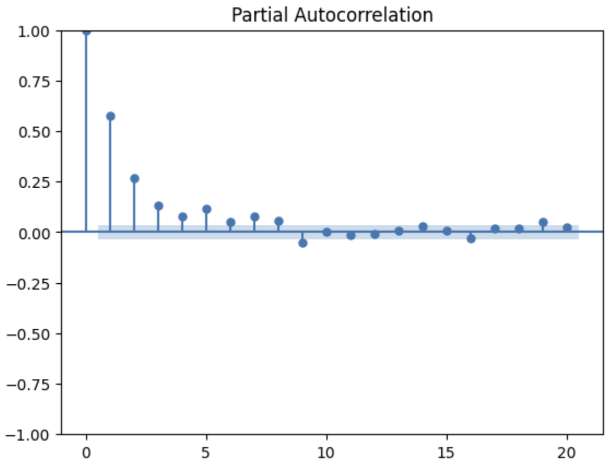

Figure 41: Partial autocorrelation plot

In both tests the non-stationarity of the historical series is highlighted which is also confirmed by reapplying the tests to the decomposed series.

In the following plot we can see the three models, ARIMA,SARIMA and Prophet fitting the validation period.

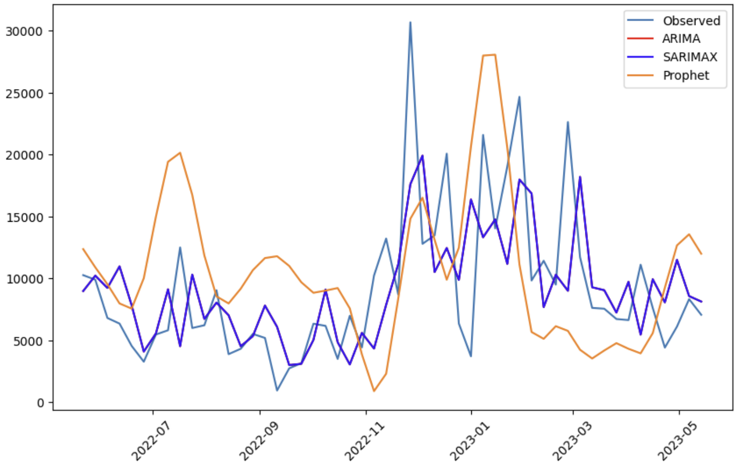

Figure 42: ARIMA, SARIMA and Prophet model fitting the validation period.

Of the three applied models the best one turns out to be ARIMA but we can notice that in this case ARIMA and SARIMA are the same model. The respective MSE values are shown in the following table.

Model	MSE
ARIMA	5004.72
SARIMA	5004.73
Prophet	7445.62

After selecting the Arima model as the most suitable for forecasting the Casual sector sales, we proceeded to compute the 95% confidence intervals to evaluate the potential range of future values.

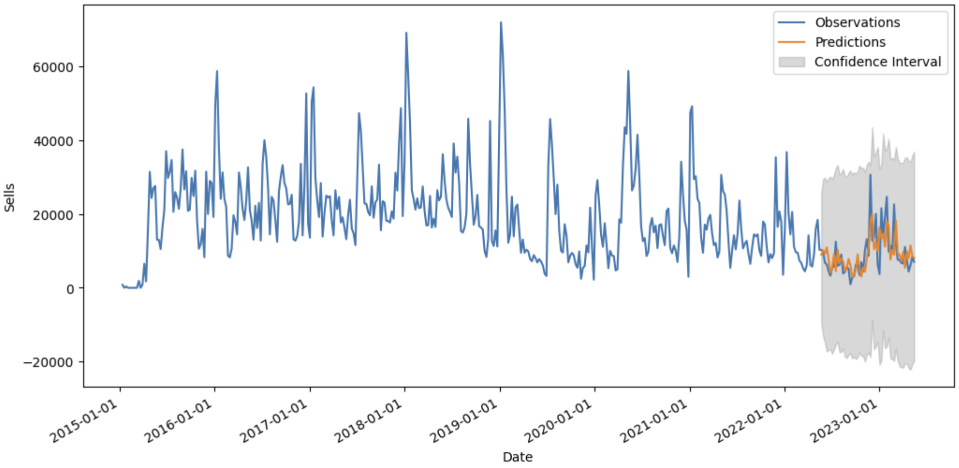

Figure 43: The best model (Arima) predictions and their relative confidence interval at 95%.

It is important to note that the intervals appear relatively wide, but this is largely influenced by the high magnitude of the sales data. Despite the broad intervals, the Arima model demonstrates quite satisfactory performance.

3.5.3 Prediction

As previously observed, sales in the Casual sector have maintained relative stability over the past 10 years without significant variations. As a result, the simulated paths may exhibit different directions, reflecting the unpredictability and intrinsic variability in sales models.

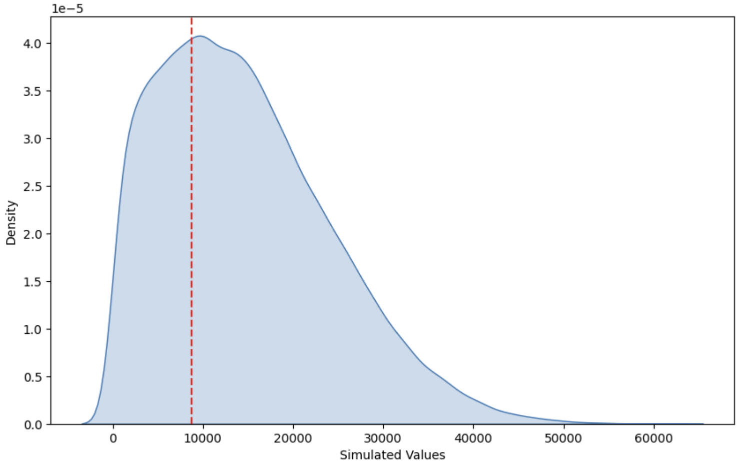

Figure 44: Distribution of daily sales for all the 1000 path.

In Figure 44, we can observe the distribution of daily sales values in the Casual sector. The average daily sales values are below 10,000, and the distribution shows, as seen before, evident asymmetry due to numerous days with zero sales. The range of values varies from 0 to a maximum of 60,000.

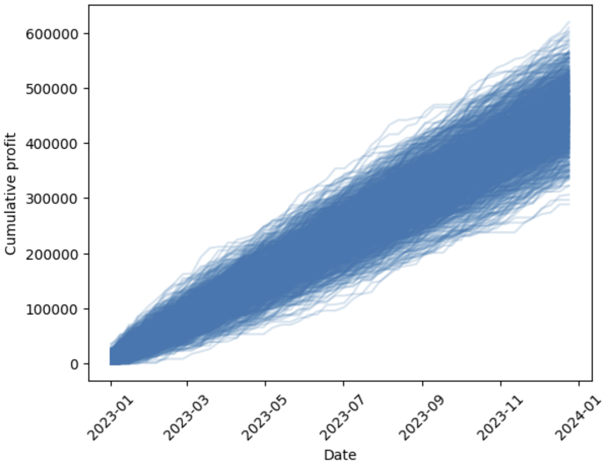

Figure 45: Cumulative curve for the 1000 simulated path.

Moving to Figure 45, we examine the cumulative distribution for the year 2023 in the Casual sector, based on the 1000 different simulations conducted. Most paths indicate cumulative earnings between 300,000 and 600,000. This suggests that cumulative sales for 2023 are expected to be similar to the previous year.

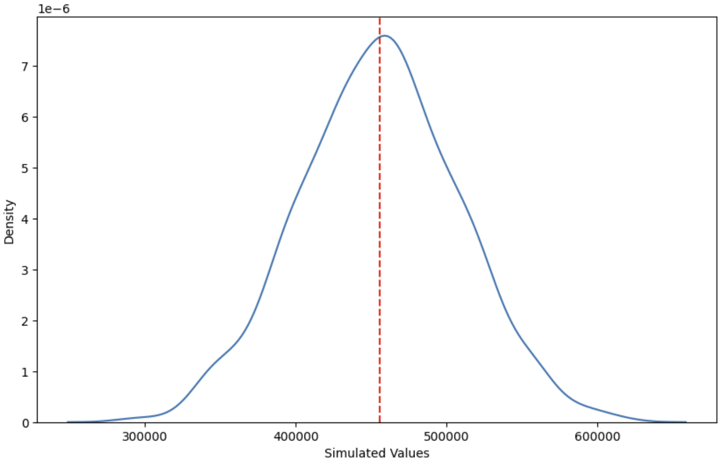

Figure 46: Distribution of the cumulative value at the end of the future year for 1000 simulated path.

Examining Figure 46, we can observe the density distribution of values across different paths in the Casual sector. It is evident that the values are concentrated around the mean value of 450,000. As previously mentioned, stability is expected in the Casual product sector, without significant increases or decreases.

3.6 Skiing

3.6.1 Exploratory Analyses

Now let's focus on the final sector under analysis: Skiing. Following the same methodology as before, we will examine the time series of sales using moving average smoothing.

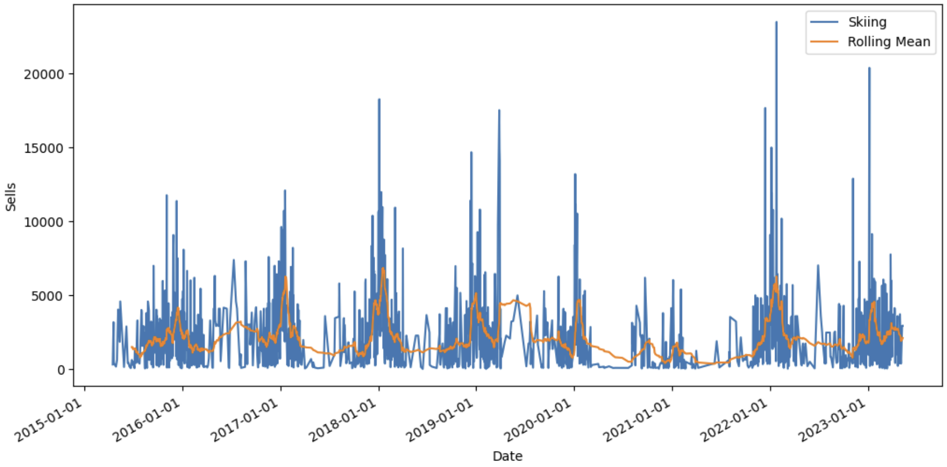

Figure 47: Skiing time series with the rolling mean smoothness.

Figure 47 presents the time series of Skiing sales, smoothed with moving average. The series does not exhibit a distinct upward or downward trend. However, there are noticeable peaks in the sales, with the exception of the year 2021. This anomaly could be attributed to the unfavorable Skiing conditions caused by the Covid-19 pandemic. Consequently, we can infer the presence of seasonality in the analyzed series.

To gain further insights, let's proceed with the representation of the data aggregated on a weekly basis.

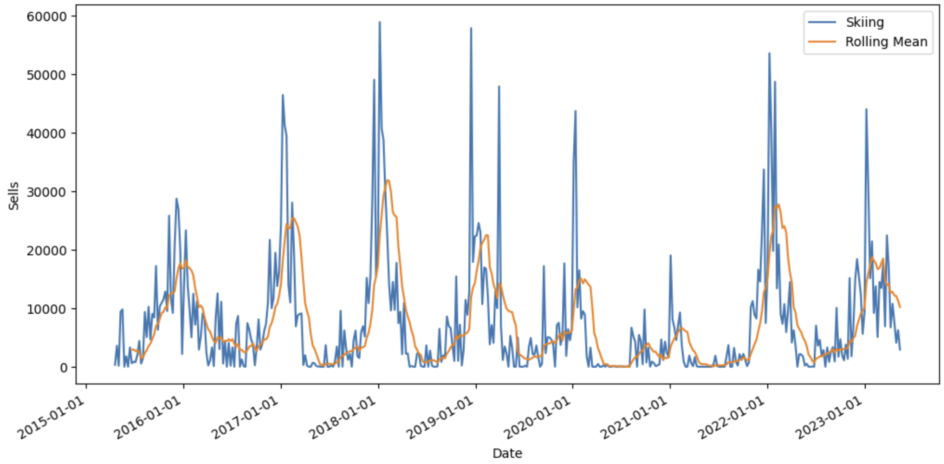

Figure 48: Skiing time series aggregated by week with the rolling mean smoothness.

Figure 48 illustrates the aggregated weekly trend for Skiing sales. The series continues to display peaks, indicating fluctuations in sales over time. However, it is worth noting that the year 2021 shows a significant deviation from this pattern, likely due to the impact of the Covid-19 crisis. Despite this, the presence of seasonality remains apparent.

Next, we will plot the data aggregated by month and by year to gain a comprehensive understanding of the Skiing sector's sales pattern.

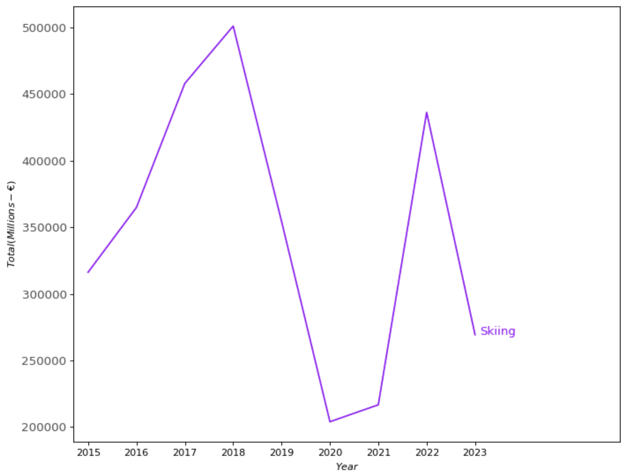

Figure 49: Cumulative gain among different years.

Figure 49 showcases the aggregated annual sales. As previously mentioned, there is a noticeable decline in sales from 2019 to 2021, followed by a strong rebound in 2022. This suggests that the Covid-19 emergency had a significant disruptive effect on a sector that had been experiencing steady growth since 2015.

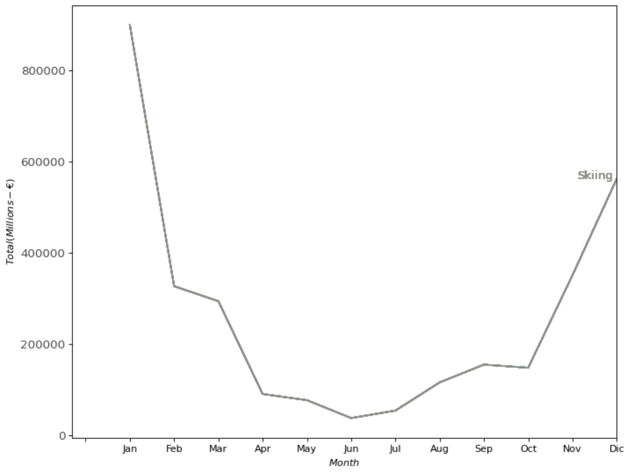

Figure 50: Cumulative gain for moth across different years.

Moving on to Figure 50, we observe the sales aggregated by month. Skiing sales exhibit a decreasing trend from February to August, with a notable surge during the winter season. This aligns with our expectations, considering the nature of Skiing as a winter sport. Let's delve deeper by analyzing the aggregated data divided by month and by year.

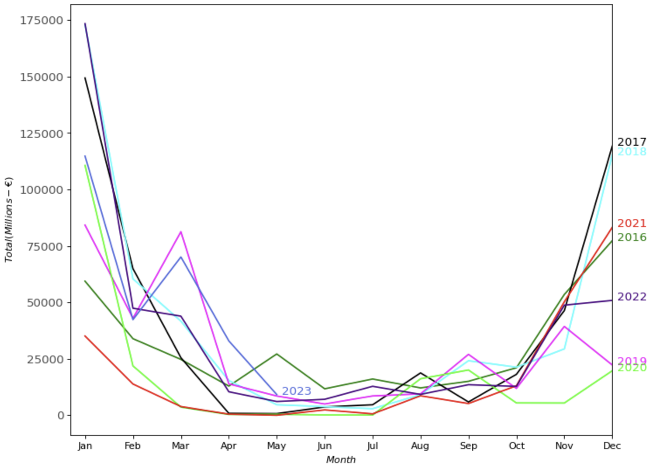

Figure 51: Cumulative gain for moth plotted for every year.

Figure 51 provides a detailed representation of the sales pattern in the Skiing sector. Excluding 2019 and 2020, the time series consistently demonstrates a similar pattern across all years. There are lower sales during the spring and summer months, followed by a robust recovery during the autumn and winter seasons. This information is valuable for marketing strategies and model development. In this regard, further detailed analyses will be conducted to refine our understanding of this sector.

3.6.2 Forecasting

Unlike the other sectors, for the Skiing sector we can see from the descriptive plots the presence of seasonal patterns, with peaks between the months of December and January months characterized by the Skiing season. This non stationary trend is less evident I the partial Autocorrelation Plot below, but we can see that some autocorrelations have non-zero value.

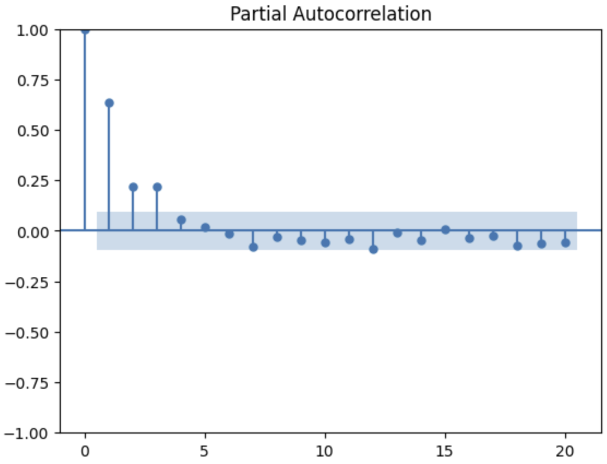

Figure 52: Partial autocorrelation plot

Despite the near certainty that this is a seasonal series the tests do not confirm this hypothesis. Despite this we expect a better performance of the SARIMA and Prophet models than the ARIMA model.

On the following plot we can see the performance of each model on the validation period.

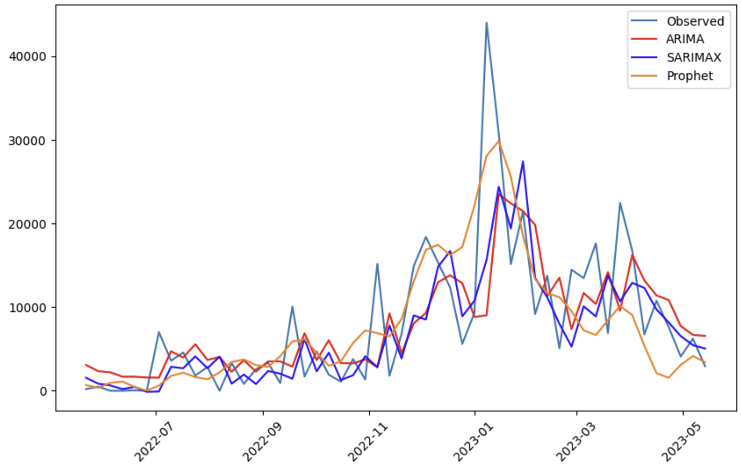

Figure 53: ARIMA, SARIMA and Prophet model fitting the validation period.

Among the three models, the one with the best performance turns out to be Prophet, with an MSE of 5418.05, but we can see that SARIMA also reports an MSE of 6075.88, which is significantly lower than the value reposted by ARIMA. The MSE results are in the following table.

Model	MSE
ARIMA	6986.86
SARIMA	6075.88
Prophet	5418.05

After selecting the Prophet model as the most suitable for forecasting the Sci sector sales, we proceeded to compute the 95% confidence intervals to evaluate the potential range of future values.

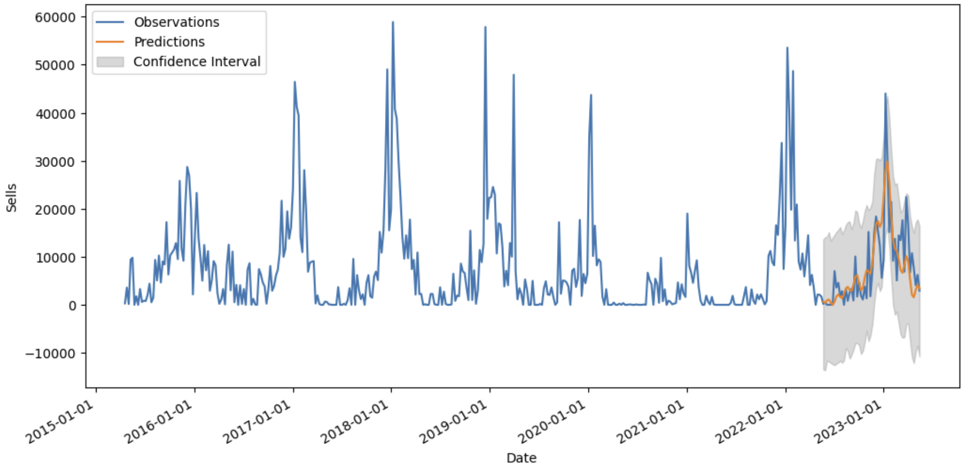

Figure 54: The best model (Prophet) predictions and their relative confidence interval at 95%.

It is important to note that the intervals appear relatively wide, but this is largely influenced by the high magnitude of the sales data. Despite the broad intervals, the Prophet model demonstrates satisfactory performance and provides valuable insights for future projections.

3.6.3 Prediction

As observed in the previous sectors, sales in the Skiing sector have remained relatively stable over the past 10 years, with low sales during the summer months and peaks between December and January. Therefore, Monte Carlo simulations may show a variety of directions, reflecting the unpredictability and intrinsic variability in sales models.

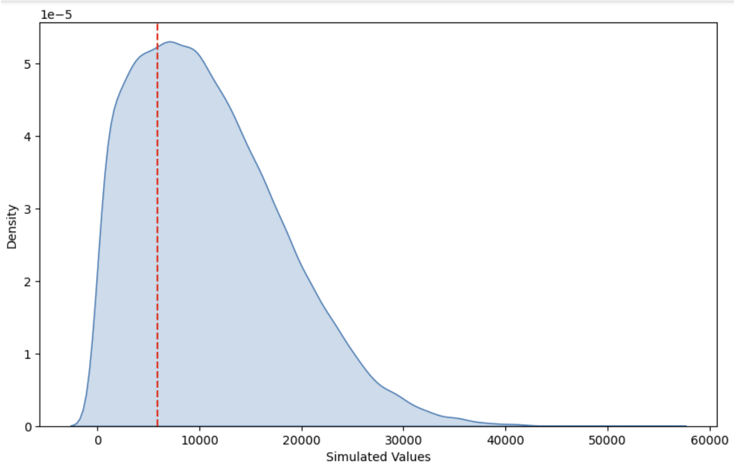

Figure 55: Distribution of daily sales for all the 1000 path.

Figure 55 illustrates the distribution of daily sales values in the Skiing sector. This chart, like others seen before, provides an idea of the realistic distribution of the predicted values. The average daily sales values remain below 10,000, while, similar to the others, the distribution's asymmetry is influenced by numerous days with zero sales. The range of values in this case ranges from 0 to a maximum of 50,000.

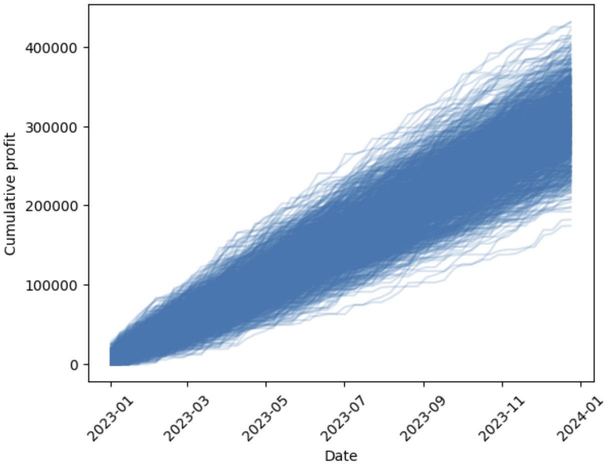

Figure 56: Cumulative curve for the 1000 simulated path.

In Figure 56, we examine the cumulative distribution for the year 2023 in the Skiing sector, obtained through 1000 different simulations. The chart shows the variety of possible outcomes for the total revenue derived from sales in the sector during the year. As mentioned earlier, a wide range of possible values, ranging from 150,000 to 450,000, can be observed. This suggests that cumulative sales for 2023 are expected to be in line with previous years, although there is no increase as seen in 2018 and 2022, likely due to the Winter Olympics.

However, if we also consider Figure 6, it seems that there is a decrease, but it should be noted that the months of November and December are still missing, which could lead to an increase in sales.

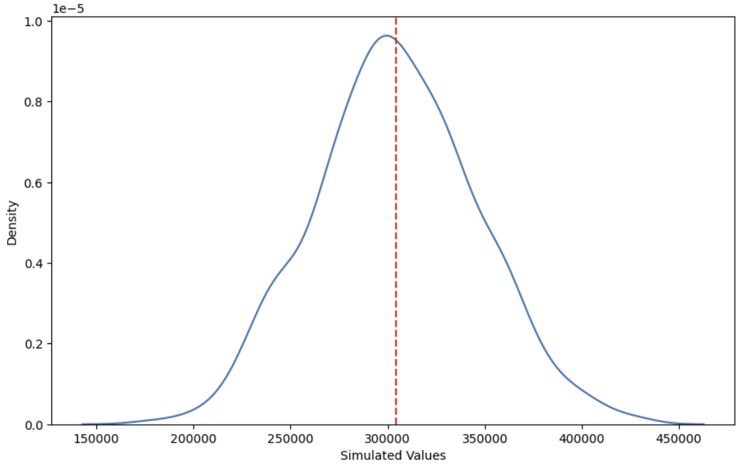

Figure 57: Distribution of the cumulative value at the end of the future year for 1000 simulated path.

Finally, in Figure 57, the density distribution of values across different paths in the Skiing sector is illustrated. It can be noted that the values are concentrated around the mean value of 300,000. As mentioned earlier, an average annual decrease in Skiing sector sales of approximately 150,000 is expected. However, compared to other years, an increase in sales is still predicted.

4 Results

Once the analysis and forecasting of the various sectors are completed, we can confidently rely on the strong results obtained. The simulation results we provided earlier are highly useful as they incorporate the inherent uncertainty associated with time series prediction. However, we now present less reliable but more easily understandable forecasting predictions that can be interpreted by non-data science professionals. These results should be considered in conjunction with the earlier simulations to gain a comprehensive and optimal understanding of the outcomes. In this section, we will showcase the predictions from the different models applied to the time series of the five sectors. By comparing these predictions, we can suggest business strategies based on future projections for the upcoming year. While these predictions are not as reliable as the simulations, they aid in understanding and guiding business decisions. It is important to consider both sets of results.

We will now present the plot of the time series predictions for the five top sectors, spanning from May 2023 to May 2024.

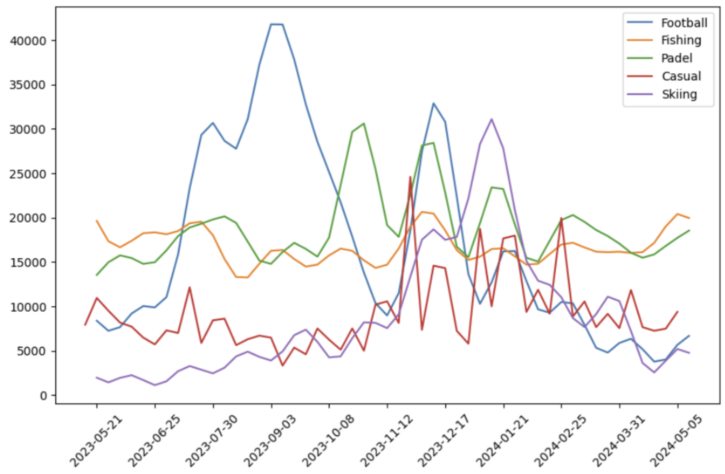

Figure 58: Prediction on the future year for the five top sectors.

From the figure, we can observe that the Football sector is projected to be highly performing during the summer to winter period. Consequently, it may be necessary for the company to increase inventory for this sector during that time. Additionally, the sales for Football, after experiencing a significant increase in 2023, are expected to decline in 2024. Therefore, investing in advertising to sustain the positive trend predicted for 2023 in the subsequent year could be beneficial. Fishing and Padel sales are anticipated to remain consistently strong compared to the other sectors, with occasional peaks. Hence, investing more in these two sectors would be advisable, considering their stable sales performance. Moreover, winter sales are projected to experience a substantial increase during the expected winter season, while in other periods, sales will be relatively low (even lower than previous years). As a result, the company could consider offering discounts on Skiing articles during non-winter periods. Lastly, the Casual sector is predicted to be volatile but relatively consistent throughout the year.

These results will assist the sales manager in making informed business decisions. However, it is important to note that the insights presented here are only suggestions, and they can be further refined by specialized professionals. The aim of this project is to provide results that effectively contribute to informed business decisions and sales strategies, ultimately leading to improved and consistent sales performance.

5 Conclusions

The analysis of the provided data has provided valuable insights for analyzing past marketing strategies and shaping future development. Through a preliminary exploratory analysis, we examined the trends of different sectors over the years, identifying Padel and Ski as the most promising sectors. Padel, in particular, has shown strong growth potential, with an upward sales trajectory since 2022. While the lack of extensive sales history limited our analysis, the data obtained so far suggests that the series exhibits stationarity. Looking ahead, we anticipate a slight decline in the high performance that Padel achieved in 2022, aligning the total gains with those of 2021. To sustain the positive trend, implementing an advertisement campaign could be beneficial in maintaining sales momentum and avoiding any potential loss of gains. The Skiing sector, is now experiencing a robust recovery. It has the potential to reclaim its position as a leading sector in e-commerce sales, particularly during the autumn/winter period. We identified a seasonal variable in the series, indicating the need for a tailored marketing strategy for the summer period. For instance, offering discounts on skiing articles during the off-season when demand is typically lower. Football, it is predicted to lose popularity in the coming years. To counteract this negative projection, implementing a campaign to sustain interest in Football could be a viable strategy. For the Casual sector, no particular seasonal component was observed. Its sales are expected to remain constant without significant fluctuations. Consequently, no specific strategy was developed for this sector. The Fishing sector emerged as the second-highest sector in terms of e-commerce sales. Similar to Padel, it does not exhibit a clear seasonal factor. Therefore, investing more in the Fishing sector, alongside Padel, would be advisable considering their consistently strong sales performance.

In conclusion, based on the analysis of the various sectors, we recommend increasing inventory for the Football and Padel sectors during their respective peak seasons. Implementing targeted advertising campaigns can help sustain positive trends in Football and Padel, while offering discounts on Skiing articles during non-winter periods can boost sales in that sector. Additionally, investing in the Fishing sector, which demonstrates steady sales performance, would also be beneficial. By considering these recommendations alongside the predictions from our simulation and forecasting models, the company can make informed business decisions and develop effective sales strategies for future success.