Data Science Lab: E-Commerce Sales

Baghrous Tariq 904027
Balbin Canchanya Gianni Eduard 901609
Barletta Aldo 897742

Master's Degree in Data Science
Università degli Studi di Milano-Bicocca
2022/2023 Academic Year

Contents

1 Introduction

In the realm of time series analysis and predictive modeling, the exploration and understanding of historical data play a crucial role in extracting meaningful insights and making informed decisions. This project is dedicated to test a diverse range of predictive models available for time series analysis in order to effectively estimate the sales generated by a set of economic sectors within an e-commerce platform.

The project was conducted by considering the three sectors with the highest availability of observations. However, the models utilized can also be applied to other branches, thereby providing a comprehensive overview of the entire e-commerce landscape.

The dataset provided underwent a preparation procedure, ensuring its efficacy and optimization in alignment with the project's goals. Indeed The predictive models we employed require a specific type of data modeling to utilize them effectively and generate accurate forecasts.

Following the data preprocessing, the project shifted its focus towards testing the following predictive models: ARIMA, PROPHET and TBATS. The analysis period selected encompass the timeframe from January 1^{st}, 2016 to May 12^{th}, 2023, while the forecast projected values up to December 31^{st}, 2023.

Lastly, the performance of the forecasting models is assested by the computation of key evaluation metrics such as RMSE, MAE, and MAPE, providing a comprehensive understanding of how to interpret and compare models with each other effectively.

2 Data Exploration

The dataset we are going to use is a "*comma-separated values*" file that contains 26615 observations and three attributes:

- **Data:** the observation date in DD/MM/YYYY format
- **Totale:** total amount of revenue in euros on that day
- **Settore:** e-commerce sector of the corresponding revenue

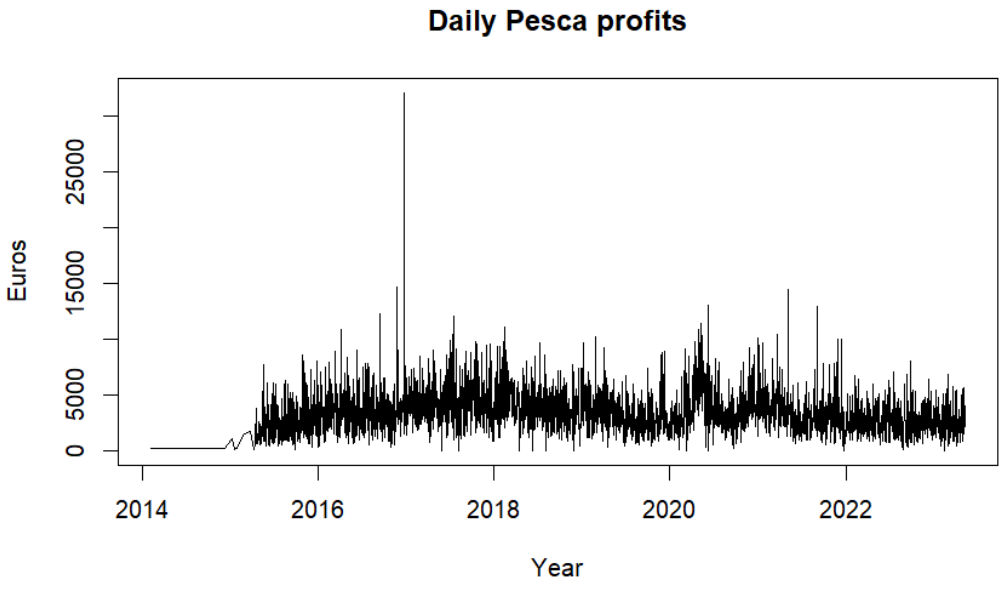
After uploading the file, we noticed that there were observations with empty **Data** attribute, so we removed them for the sake of our project, therefore reaching 26585 observations.

We created then a concise and informative summary of the dataset, detailing the number of rows for each unique **Settore** through a loop that systematically iterated over each sector calculating the number of rows in the dataset corresponding to that sector.

After doing this computation, we found out that the first three sectors by number of sales are respectively **Pesca**, **Calcio** and **Casual**. Now let's see how the data are distributed inside these sectors.

2.1 Pesca

The feature **Pesca** has 2966 rows, implying that there have been 2966 sales in this sector in a span of ten years. Below, we see now how these sales are distributed throughout the years:



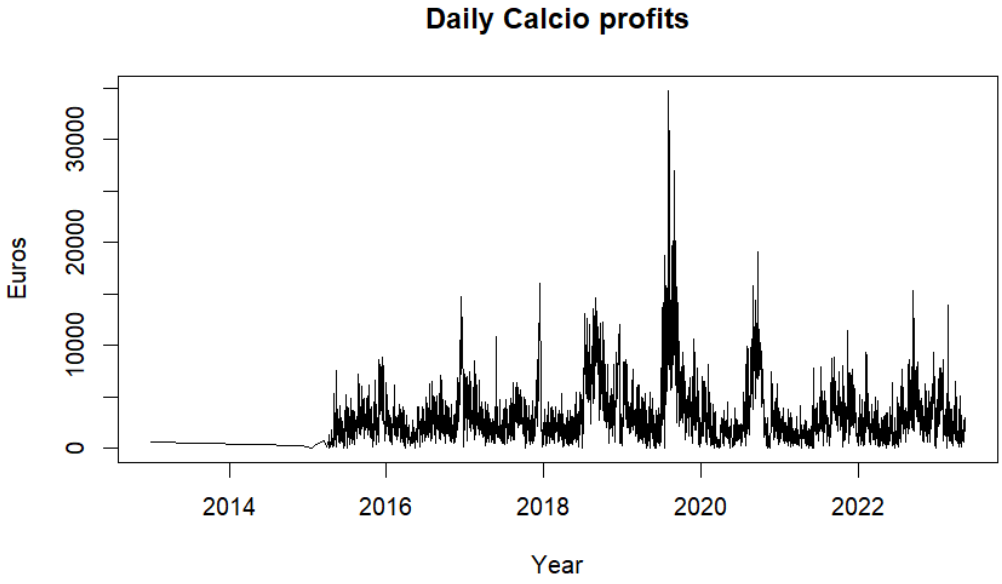
With the exception of a few outliers, such as the one with a sales value of €32049.832 on December 22, 2016, which coincides with a critical time of year, just three days before Christmas, the sales data appears relatively stable around the values of €5000 over the years. However, one notable anomaly stands out: a significant spike in the latter months of

2020. This spike suggests a surge in the sale of fishing-related tools following the outbreak of the Covid-19 pandemic.

This trend is likely to recur across other sectors, as we will see later on. What it is still worth to mention is that, even though there are two instances in which the sales have been done in the year 2014, the true gathering of the data starts from January 10, 2015, therefore we have a visible "gap" at the beginning of the plot. This observation underscores the impact of external events, such as the pandemic, on consumer behavior and market dynamics on the sales of every e-commerce sector.

2.2 Calcio

The sector **Calcio** has 2942 observations, meaning that there have been 2942 sales in the football sector between 2014 and 2023. As previously, we are going to see now the distribution of data in this particular sector:

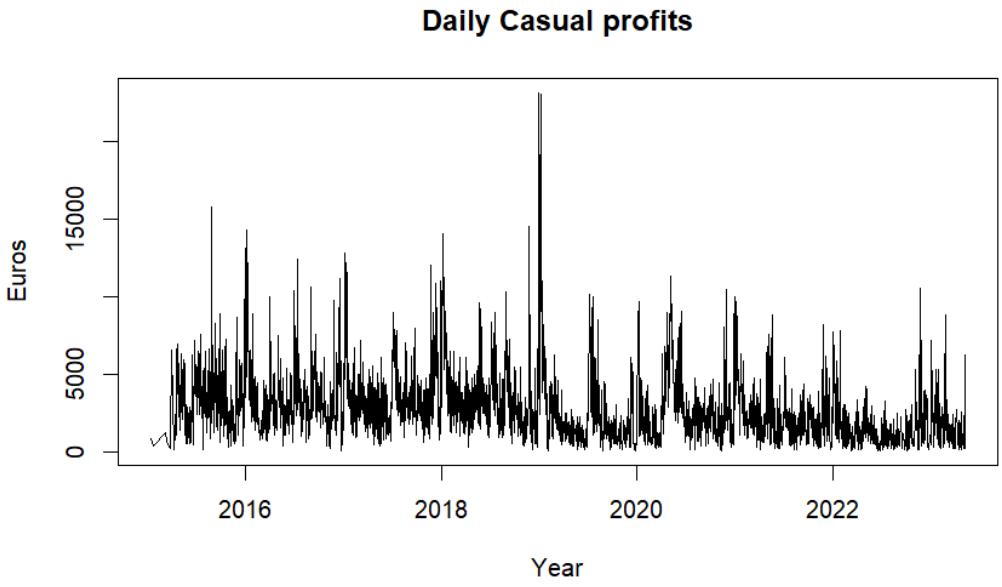


As evident from the data, we can observe distinct peaks at various points throughout the years. Notably, these spikes occur predominantly in the months of August and December, with daily purchases surpassing the €10,000 mark. The August outliers in 2019 are particularly remarkable, where sales exceeded €30,000.

These anomalies can be attributed to specific seasonal trends: the December peaks align with the tradition of purchasing gifts for Christmas, as we observed in the fishing sector earlier, while the August spikes are associated with the commencement of the football season, prompting higher spending in preparation for the games. Moreover, in the year of 2014 we have only two observations, therefore we have a gap between 2014 and 2015, like it happened in the **Pesca** sector.

2.3 Casual

The sector **Casual** has 2879 rows, implying that there have been 2879 sales in this sector in a span of eight years (2015-2023). As per usual, we see now how these sales are distributed throughout the years:



The intriguing aspect within the casual e-commerce sector is the presence of seemingly random spikes in sales patterns. These spikes are not easily attributable to predictable factors like seasons or specific marketing campaigns. Instead, they appear to occur unpredictably, defying conventional patterns.

One plausible explanation for this phenomenon could be the influence of various external and internal factors that introduce volatility into the casual e-commerce market, like for instance the consumer behaviour, which are driven by individual preferences and trends that can change rapidly, or external events, like awards shows, fashion weeks, or even unexpected global events can influence fashion choices.

What is also different in this data distribution in respect to the oter ones is that there are no data available for the year 2014, therefore we do not have the aforementioned problem of the "gap" inside the plot.

3 Models

As introduced, the forecasting models we are going to adopt are ARIMA, TBATS and PROPHET, each offering distinct approaches to modeling time series data. The selection of these prediction models was evaluated considering a way to inherently handle missing data without requiring explicit imputation steps. Thus, we are now to provide an overview of these models and their applications.

3.1 AutoRegressive Integrated Moving Average (ARIMA) Model

The AutoRegressive Integrated Moving Average (ARIMA) model stands as a versatile and powerful framework for comprehending and forecasting complex data patterns. While AutoRegressive (AR) and Integrated Moving Average (IMA) models tackle specific aspects of time series data, the ARIMA model amalgamates these components to provide a holistic approach to modeling and forecasting.

The ARIMA model of order ARIMA(p, d, q) can be described as follows:

$$Y_t = \mu + \phi_1 Y_{t-1} - \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$$

Here, μ represents the mean of the series, ϕ_i are the coefficients of the autoregressive component, ϵ_t is the differenced series, and θ_i are the coefficients of the moving average component. The ARIMA model ensures that the expected value of the differenced series ϵ_t is zero.

Several key implications emerge from the ARIMA model:

1. **Combination of AR and IMA:** The ARIMA model seamlessly integrates both AutoRegressive (AR) and Integrated Moving Average (IMA) components. The AR component captures lagged relationships, while the IMA component leverages differencing to achieve stationarity.

2. **Differencing for Stationarity:** Similar to the IMA model, the ARIMA model incorporates differencing, specifically d times, to render the time series stationary. Stationarity is a fundamental prerequisite for accurate modeling and forecasting.

3. **Flexibility in Complexity:** The ARIMA model allows for a range of complexity, as indicated by the orders p, d, and q. The choice of these orders depends on the specific characteristics of the time series under examination.

The ARIMA model's flexibility and adaptability make it a valuable tool for capturing a wide array of temporal patterns and dependencies in data. It serves as a comprehensive framework that combines the strengths of both AR and IMA models.

For an ARIMA(p, d, q) model to be effective, it must successfully balance the autoregressive, differencing, and moving average components. The selection of the orders (p, d, and q) is a critical decision and should be based on a thorough understanding of the data's behavior.

In summary, the AutoRegressive Integrated Moving Average (ARIMA) model represents a robust approach to time series analysis and forecasting. Its adaptability and ability to capture complex patterns make it a fundamental tool in the field. Understanding how to tailor the model's components to specific datasets is essential for harnessing the full potential of ARIMA in forecasting and analysis.

3.2 Trend-Based Seasonal Decomposition of Time Series (TBATS) Model

The TBATS (Trend-Based Seasonal Decomposition of Time Series) model distinguishes itself as a sophisticated and adaptable method for forecasting time series. In contrast to conventional models, TBATS goes further by integrating multiple seasonal components, enabling it to effectively capture complex patterns frequently present in real-world data.

The TBATS model can be expressed as:

$$Y_t = \mu + \text{Trend Component} + \text{Seasonal Components} + \text{Residuals}$$

Breaking down the key elements of the TBATS model:

1. **Trend Component:** Expressing the fundamental trend within the time series, this component has the capability to capture linear, exponential, or damped trend patterns. Such flexibility empowers the model to adjust to different directional trends exhibited in the data.
2. **Seasonal Components:** TBATS is designed to handle multiple seasonal components, rendering it highly suitable for time series data characterized by diverse seasonal patterns. These components may encompass weekly, monthly, or any recurring patterns evident in the dataset.
3. **Residuals:** TBATS addresses unexplained variations in the time series by working to minimize residuals through the effective modeling of both the trend and seasonal components.

TBATS offers several advantages, enhancing its utility in time series analysis:

- **Handling Multiple Seasonalities:** TBATS excels in situations where time series data exhibits multiple seasonal patterns, providing greater flexibility compared to traditional models.
- **Automatic Parameter Selection:** The model autonomously determines suitable parameters, including the number of harmonics for each seasonal component. This reduces the necessity for manual tuning, thereby enhancing overall efficiency..
- **Robustness to Irregularities:** TBATS showcases resilience to irregularities in the data, such as missing values or outliers. This characteristic positions it as a well-suited choice for handling real-world datasets with diverse characteristics.

Implementing the TBATS model typically involves using optimization techniques to estimate parameters that best fit the observed time series data. The resulting model can then be employed for forecasting future values based on the learned patterns.

In summary, the Trend-Based Seasonal Decomposition of Time Series (TBATS) model offers a powerful and adaptable approach for time series forecasting. Its ability to handle multiple seasonalities and adjust to various trend patterns makes it a valuable tool for capturing complex temporal dependencies in diverse datasets. Understanding the components and strengths of TBATS is crucial for effective application in forecasting and analysis.

3.3 PROPHET

Prophet, a forecasting tool created by Facebook, is an open-source solution made for the precise forecasting of time series data. Its strength shines when dealing with datasets

characterized by recurring seasonal patterns, holiday effects, and unique events. Let's delve into the core features of Prophet:

- **Seasonality Modeling**: Prophet is excellent at handling time series data that have multiple seasonal patterns. For example, sales data for a product might exhibit daily, weekly, and yearly seasonality (e.g., higher sales on weekends, higher sales during holiday seasons). It can automatically detect and model these seasonal patterns without requiring manual intervention.

- **Holidays and Special Events**: Prophet allows you to explicitly include holidays and special events that can significantly impact your time series data. This makes it a powerful tool for forecasting in scenarios where holidays or other events have a noticeable effect on the data.

- **Robustness to Outliers and Missing Data**: Prophet is designed to handle outliers and missing data points gracefully. It uses a procedure that makes it less sensitive to extreme values, making it a robust choice for real-world datasets that might have irregularities.

- **Automatic Trend Change Detection**: Prophet can automatically detect and model abrupt changes in a time series. For example, if a new marketing campaign significantly impacts sales, Prophet can adapt to this change without manual intervention.

Overall, Prophet is a powerful tool for time series forecasting, especially when dealing with data that has multiple seasonal patterns or events that influence the observations. It's user-friendly and provides a robust framework for generating accurate forecasts.

4 Data Analysis

After having a look at the three sectors of our interest, it is time to analyze them and preparing the dataframes for exploiting the models explained in the paragraph 3. But first, we have to introduce three tools useful to find stationarity and stagionality patterns: the **Augmented Dickey-Fuller (ADF) Test**, the **Autocorrelation Function, (ACF)** and the **Partial Autocorrelation Function (PACF)**

- The **Augmented Dickey-Fuller (ADF) test** is a statistical test used to determine whether a given time series is stationary or non-stationary. A stationary time series is one whose statistical properties, such as mean, variance, and autocorrelation, remain constant over time. Non-stationary time series, on the other hand, exhibit trends, seasonality, or other patterns that evolve over time. The ADF test produces a test statistic, with a more negative value providing stronger evidence against the null hypothesis. Additionally, it calculates a p-value, and if the p-value falls below a predetermined significance level (e.g., 0.05), the null hypothesis is rejected, indicating stationarity. Conversely, a higher p-value suggests non-stationarity.

- The **Autocorrelation Function (ACF)** measures the correlation between a time series and its own lagged values. It is calculated for various lag values, and the resulting ACF plot shows how correlated the time series is with itself at different lags. ACF helps in identifying the presence of any systematic patterns, cycles, or trends in the data.

- The **Partial AutoCorrelation Function (PACF)** measures the correlation between a time series and its own lagged values while controlling the effects of the other lags. It is useful in identifying the direct relationship between the current observation and its past observations, excluding the influence of intermediate lags. PACF is particularly helpful in determining the order of an autoregressive (AR) model, which is a type of time series model.

4.1 Pesca

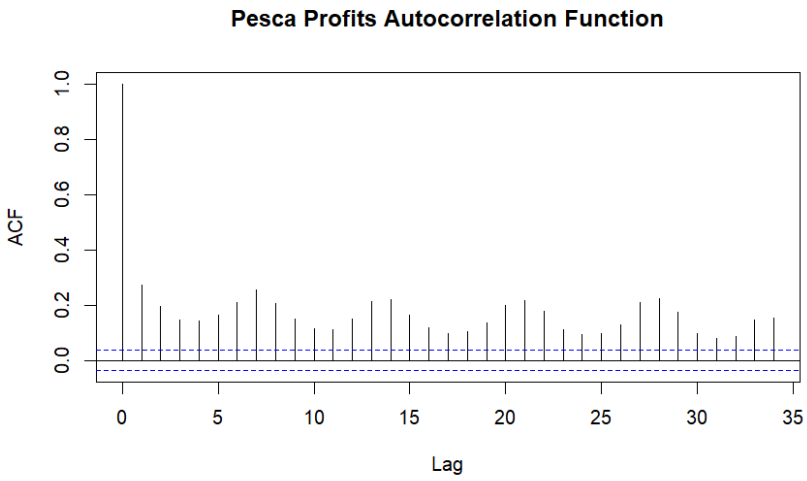
4.1.1 Data manipulation

The dataset was examined to assess the number of observations per year. First of all duplicates were identified and combined, consolidating redundant information. Then, entries for the years 2014 and 2015 were excluded due to lack of sufficient data. The final dataset, free from duplicates and specified years, has 2687 observations and is ready for the next step.

4.1.2 Stationarity and stagionality

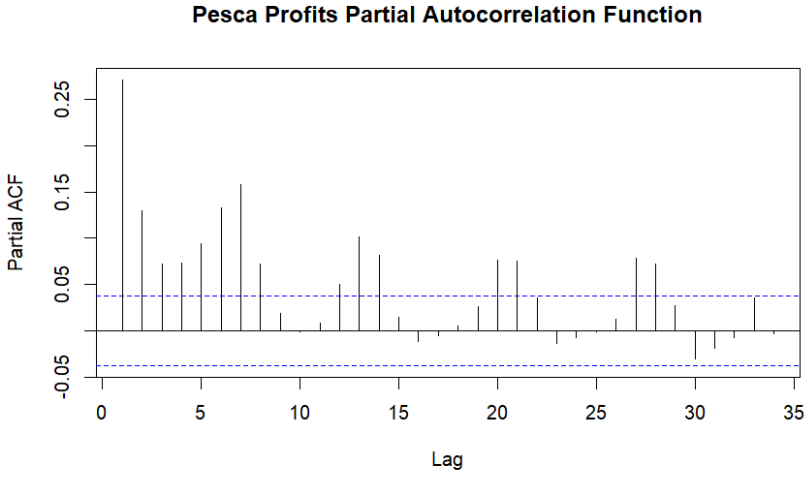
The Augmented Dickey-Fuller (ADF) test was conducted to assess stationarity in the profits data. The test produced a Dickey-Fuller statistic of **-7.931** with a lag order of **13** and a p-value of **0.01**. The obtained p-value, being below the conventional significance level of 0.05, leads to the rejection of the null hypothesis, suggesting that the time series data is stationary. This implies that the 'Pesca' profits exhibit a stable pattern over time, making them amenable to time-series analysis and forecasting models.

Autocorrelation and partial autocorrelation functions were visualized to identify potential seasonality patterns:



Positive Correlation at Lag 1: The positive correlation observed at lag 1 (0.27) suggests a potential trend in the data. An interesting value at lag 1 implies a strong dependence of current time series values on their preceding values.

Significant Correlations at Multiple Lags: Notable correlations were found at lag 2 (0.19) and lag 7 (0.21). This indicates the possibility of a recurring seasonality or cycle at specific intervals within the dataset.



Significant correlations emerge at various lag intervals, hinting at the potential existence of seasonality or a repeating cycle. However, after lag 6, these correlations show a diminishing trend, indicating a weaker dependency. PACF analysis points towards a trend at lag 1 and raises the possibility of seasonality at multiple lags.

4.2 Calcio

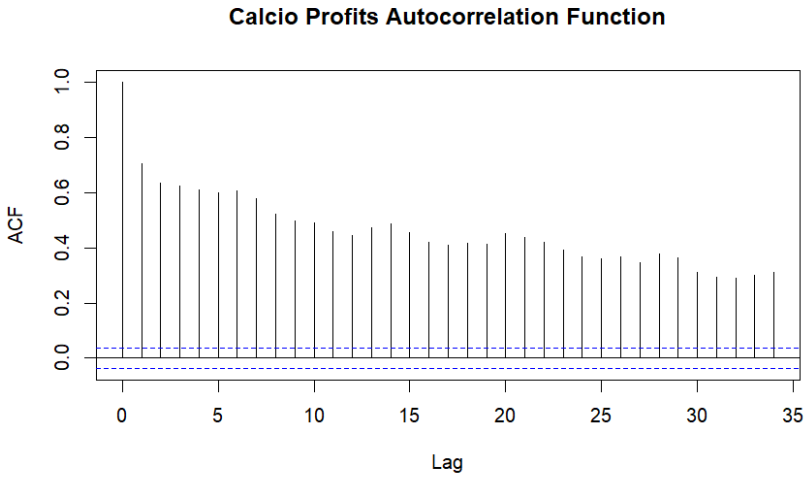
4.2.1 Data manipulation

Similar to the analysis conducted for the 'Pesca' dataset, an evaluation of the 'Calcio' dataset was undertaken to determine the annual count of observations. The process involved identifying and merging duplicates to streamline redundant information. Subsequently, data entries corresponding to the years 2013, 2014 and 2015 were excluded due to insufficient data availability. The resulting dataset, devoid of duplicates and data from the specified years, comprises 2664 observations and is now prepared for the subsequent analysis.

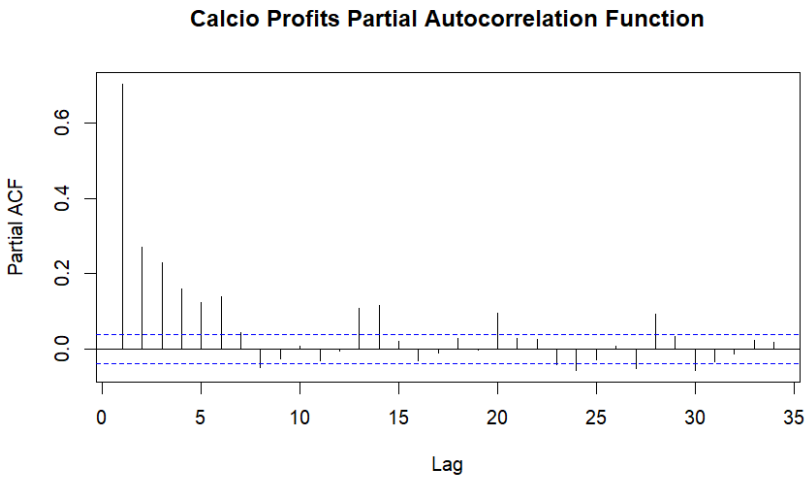
4.2.2 Stationarity and stagionality

The Augmented Dickey-Fuller Test yielded a Dickey-Fuller statistic of -5.5196 with a lag order of 13 and a p-value of 0.01. The p-value, being less than the conventional significance level of 0.05, leads to the rejection of the null hypothesis, indicating that the time series data is stationary. This implies that the 'Calcio' profits exhibit a stable pattern over time, making them suitable for time-series analysis and forecasting models.

Autocorrelation and partial autocorrelation functions were visualized to identify potential seasonality patterns:



It appears that the correlation decreases as the lag increases. The autocorrelation function (ACF) exhibits significant values at early lags but diminishes as we move away from zero. This pattern may suggest the presence of a potential trend or seasonality in the time series. The persistence of significant values at regular intervals could indicate the existence of seasonality in the time series.



There is a trend in the initial observations of the time series, but this trend diminishes in significance as we move forward in time. There appears to be a significant correlation with the previous day and the day before that, suggesting a possible daily trend or a seasonal structure at two-day intervals.

In summary:

- Trend: The Dickey-Fuller test suggests that the time series is stationary, indicating the absence of a significant trend.
- Seasonality: ACF and PACF show evidence of correlation in the first two lags, suggesting a potential seasonality in the time series.

4.3 Casual

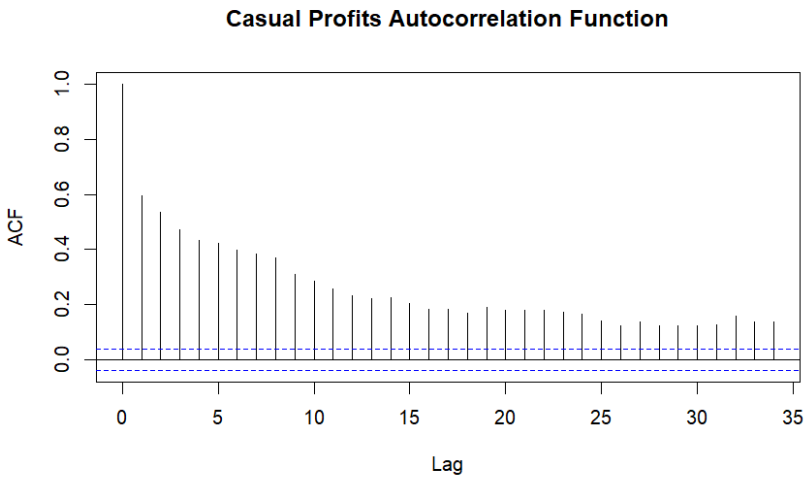
4.3.1 Data manipulation

Similar to previus datasets, also this one underwent an analysis to determine the annual count of observations. Initially, duplicate entries were identified and merged, then data entries pertaining to the year 2015 were excluded due to insufficient data. The resulting dataset, devoid of duplicates and entries from the specified years, comprises 2606 observations and is now prepared for the subsequent analysis.

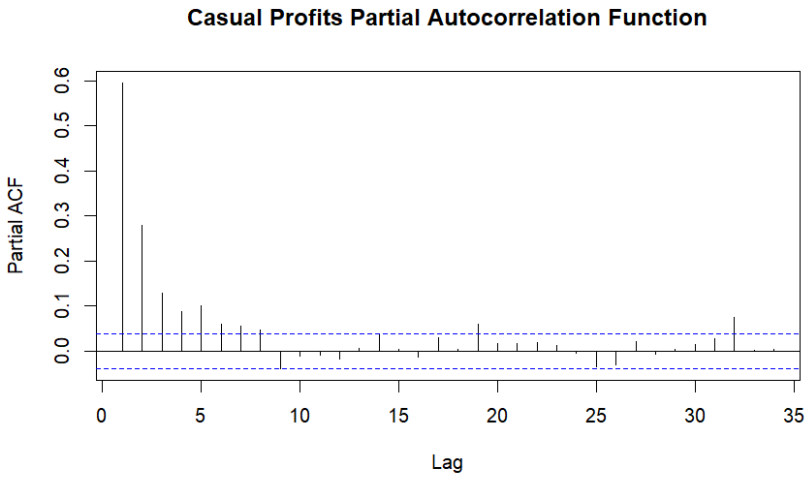
4.3.2 Stationarity and stagionality

The Augmented Dickey-Fuller test results indicate a test statistic (Dickey-Fuller) of -9.4026, a lag order of 13, and a p-value of 0.01. With a p-value less than the significance level (0.05), there is sufficient evidence to reject the null hypothesis. Therefore, based on the low p-value and the negative test statistic, we can conclude that the dataset is stationary.

Autocorrelation and partial autocorrelation functions were visualized to identify potential seasonality patterns:



The initial lag is 0.60, which could suggest a strong correlation with the previous lag, indicating a trend. Lags at periods 11 and 22 exhibit significant values, suggesting a potential seasonal cycle of 11 or 22 periods.



From the PACF, it appears that there is a significant correlation at least up to the first lag, suggesting the presence of a possible trend. The PACF also hints at the possibility of seasonality or other significant patterns at intermediate lags. Performing differencing is not recommended. If differencing leads to a time series that looks like white noise, it could indicate that the original series was already stationary or that the applied differencing model is too strong.

5 Results

5.1 Models Evaluation

In order to evaluate the models, we are going to use four indices which play a crucial role in assessing the performance of predictive models. They are the following:

- The **Root Mean Squared Error (RMSE)** measures the average errors between predicted and actual values. A lower RMSE indicates better model performance, because it means smaller and more consistent prediction errors.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- The **Percentage Root Mean Squared Error (%RMSE)** is a variation of RMSE that expresses the error as a percentage of the average actual values. This metric is useful for understanding the magnitude of errors in the context of the overall variability in the data.

$$\%RMSE = \frac{\text{RMSE}}{\bar{Y}} \times 100$$

- The **Mean Absolute Percentage Error (MAPE)** is the average percentage difference between predicted and actual values. It provides a straightforward interpretation as it represents the average percentage error relative to the actual values.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

- The **Median Absolute Percentage Error (MdAPE)** is a robust alternative to MAPE that calculates the median of the absolute percentage errors. Unlike MAPE, it is less sensitive to extreme values, making it a suitable metric for datasets with outliers.

$$\text{MdAPE} = \text{Median}\left(\left|\frac{A_t - F_t}{A_t}\right|\right)$$

The training of predictive models was made by splitting the datasets into training and test sets following a proportion of respectively 80% - 20%. The results are reported in the following tables, for each business sector:

Pesca				
Model	RMSE	%RMSE	MAPE	MdAPE
ARIMA	2483.639	92.10784	124.8932	54.3886
TBATS	1439.086	53.36971	77.54023	36.64696
PROPHET	1448.853	53.73193	79.76518	38.17141

Calcio				
Model	RMSE	%RMSE	MAPE	MdAPE
ARIMA	2973.671	101.8145	273.9152	71.18257
TBATS	2241.583	76.7488	129.4034	50.51817
PROPHET	2556.560	87.53316	96.81574	74.8963

Casual				
Model	RMSE	%RMSE	MAPE	MdAPE
ARIMA	2510.836	152.6624	292.0526	88.68437
TBATS	1520.031	92.42	209.2602	65.579
PROPHET	1681.554	102.2409	216.5472	64.94955

In light of the above-displayed results, we can observe that the most effective predictive model is TBATS, capable of providing a lower forecasting error compared to other models, at least in the majority of analyzed cases. This holds true both across different sectors and various indices.

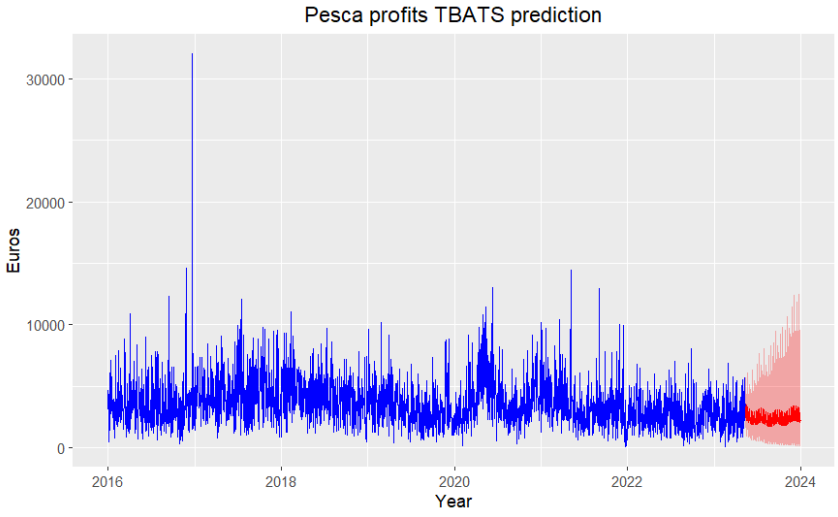
In particular, we have noted that the top-performing model in the Pesca sector is TBATS for every indices calculated. For the Calcio sector, TBATS recorded the lowest RMSE and MdAPE but PROPHET had a better MAPE. In conclusion, for the Casual sector, PROPHET was the best overall model with the lowest MAPE and MdAPE recorded.

On the other hand, the model with the least favorable performance is the ARIMA with the highest errors in every case explored.

5.2 Forecasts

5.2.1 Pesca

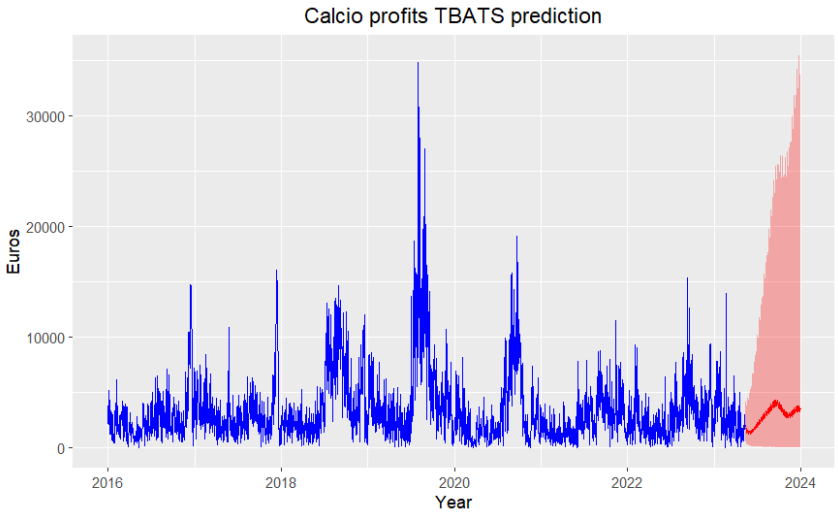
As said before, the TBATS model recorded the lowest RMSE, %RMSE, MAPE, and MdAPE for the Pesca market. The forecasting was developed by considering both a weekly and yearly seasonality while fitting the model.



The range or uncertainty of the predicted values were computed with a confidence interval of 80%.

5.2.2 Calcio

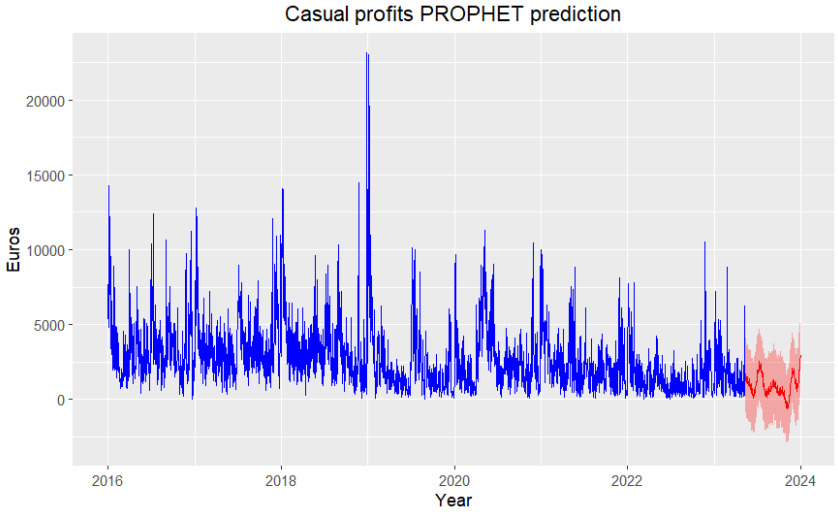
Also in this case the TBATS model recorded the overall best performance. The forecasting was developed by considering both a weekly and yearly seasonality while fitting the model.



For this sector we have also achieved good results with the PROPHET model, considering the low MAPE achieved in comparison with the other models.

5.2.3 Casual

the PROPHET model recorded the lowest MAPE and MdAPE for the Casual sector. The forecasting trend showed below appears to record a decrease in sales for the remainder of 2023.



6 Conclusions

The objective of this project was to identify the best predictive model among those analyzed to enable the analysis of sales in an e-commerce across various proposed sectors. The results obtained can be considered satisfactory.

The TBATS and PROPHET models proved particularly effective in forecasting for the Pesca, Calcio, and Casual sectors. Specifically, TBATS demonstrated superior performance in the Pesca and Calcio sectors, accurately capturing the complex dynamics of sales data. On the other hand, the PROPHET model excelled in the Casual sector, adapting better to the unique dynamics of this industry.

It is important to highlight that the chosen forecast period is extensive, and the ability to update results over time with new observations offers a dynamic approach. Reducing the temporal interval could enhance accuracy but might overlook important information.

Furthermore, extending the analysis to other sectors within e-commerce would be stimulating, deepening the understanding of the specific dynamics of each sector.

As future developments, testing other algorithms for time series prediction could be considered. This approach could provide additional insights and further improve predictions, especially in more complex scenarios or with highly dynamic data.

In conclusion, the combined use of TBATS and PROPHET models has proven to be an effective strategy, but exploring new methodologies could lead to further enhancements in e-commerce sales analysis.

References

[1] Time Series Forecasting in R - PROPHET, https://www.kaggle.com/code/rahuldhamne/time-series-forecasting-in-r-prophet

[2] TBATS model, https://www.rdocumentation.org/packages/forecast/versions/8.21.1/topics/tbats

[3] Marco Fattore. Fundamentals of time series analysis, for the working data scientist. 2022.