Predictive analysis of sports e-commerce sales: exploring trends and forecasts

Mattia Birti 897092, Samir Doghmi 897358, Gloria Longo 864579, Davide Prati 845926

Abstract
In this project a dataset will be analyzed containing all the sales of an e-commerce specialized in sports items from 12-11-2012 to 04-04-2023. The project is mainly developed in two parts: the first part consists in the integration of the data set and in its exploration, the second instead consists in the implementation of predictive models in order to forecast the future turnover of the company in question.The purpose of the analysis is to investigate the most influential patterns for online purchases by consumers and train predictive algorithms capable of predicting the future turnover of companies owning an online store.

Contents

Introduction 1
Description of the dataset . . . 1
Objective . . . 2
Structure of the report . . . 2
1 Data exploration and preprocessing 2
1.1 Overview of the dataset . . . 2
1.2 Preprocessing . . . 2
2 Data integration 3
2.1 Covid-19 . . . 4
2.2 Winter and summer markdowns and Black Fridays 4
2.3 Holidays . . . 5
3 Data visualization 5
4 Models 7
4.1 ARIMAX . . . 7
4.2 Prophet . . . 9
4.3 XgBoost . . . 10
4.4 Multivariate Linear Regression . . . 11
4.5 SARIMAX . . . 12
5 Evaluations 13
5.1 Mean Absolute Error . . . 13
5.2 Mean Squared Error . . . 13
5.3 Root Mean Squared Error . . . 13
5.4 Mean Absolute Percentage Error . . . 13
5.5 Best Models . . . 13
Conclusions and future developments 14
References 14

Introduction

Over the last decade, the concept of the online store has changed and grown a lot in our daily lives. Just think of large companies such as Amazon that invoice millions of eu ros every year through their e-commerce. With the advent of Covid-19 the concept of e-commerce quickly spread and normalized throughout society as it was the only way to shop during that era. Of course, with the reopening of brick and mortar stores, people have started shopping in those stores again but in fewer numbers than before as they have become accustomed to online stores.
This project will analyze the turnover of a sports-themed e-commerce that contains all the sales from 12-11-2012 to 04-04-2023.
During the exploration phase, the possible factors that influence the purchase in the online shop of consumers will be investigated. In the second phase, however, predictive algorithms will be implemented and analyzed to forecast future sales.

Description of the dataset

Our original dataset refers to a sport e-commerce store, and lists in rows all the orders ranging from 11-12-2012 to 04-04-2023. For each of them the dataset presents seven columns, which in detail are:

- **id** (numeric - ratio):
  Just an identifier associated to each sale and that does not add any significant information
- **marchio** (categorical – nominal):
  It represents the brand of the sold product
- **descrizione** (categorical - nominal):
  It just adds a description of the product

- **settore** (categorical - nominal):
  It contains the sector of the product it refers to, as for example its designated sport activity, like 'Running' or 'Snowboard', or more in general the product category, like 'Casual'
- **qta** (numeric - ratio):
  The number representing the quantity of the product it refers to that is sold in that particular sale
- **prezzo** (numeric - ratio):
  The price of that purchase, expressed in Euros
- **data** (numeric - interval):
  It contains the sale's date and time, expressed in the format year-month-day hour:minute:second.millisecond (YYYY-MM-DD hh:mm:ss.SSS).

Objective

Fist of all we have to specify that our project can be considered as addressed directly to e-commerce managers and intended for use by those who evaluate sales and revenues on behalf of the company, as well as scheduling and handling warehouse management. We will adopt this assumption for the entirety of the report.
Stated that, the goal of the project is to test and evaluate some predictive models and then identify the best one among them, in order to provide to the managers of the e-commerce platform an optimal model which can estimate sales in euros for each sector.
Being able to know revenue forecasts of the various sectors of the products sold by the company could in fact be useful for several purposes, and in particular it can be the basis on which to develop sales choices and strategies, such as planning when to apply or not a particular discount or promotion to a particular product or sector.

Structure of the report

This report is organized as follows:

1. **Data exploration and preprocessing**: we examine features of the original dataset, then we add some columns to the dataset, transform some features and handle particular values in order to make the dataset more suitable for analysis.
2. **Data integration**: in this section we integrate our data with information obtained from other sources, in order to enrich the dataset.
3. **Data Visualization:** this section is used just to present some graph and results that we haven't had the opportunity to show previously and that are useful to better understand the dataset potential.
4. **Models**: in this phase, we train predictive models suitable for the analysis of time series to allow us to achieve our goal.
5. **Evaluations**: in this last section we process and compare evaluation metrics on the models implemented before.

1. Data exploration and preprocessing

1.1 Overview of the dataset

Before doing anything else, we start our work by having a first glance at the original row dataset as provided us by the company.

Dataset statistics

Number of variables	7
Number of observations	378243
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	20.2 MiB
Average record size in memory	56.0 B

Figure 1. *Initial overview of the dataset*

Using Pandas Profiling Report [1], we developed a brief overview of the dataset: as we can see in the figure above, to be precise our dataset is constituted by 378 243 observations and 7 variables, as already said.
An absolutely relevant consideration is that luckily we don't have any missing values or any duplicate rows, so we don't have the issue to deal with them during the phase of preprocessing.

Variable types

Numeric	3
Categorical	3
DateTime	1

Figure 2. *Numeric, categorical and date time variables*

Analyzing variable types, we can see that the original dataset stores the values of column "data" in a datetime format.

1.2 Preprocessing

So, regarding datetime issues discovered in data exploration, we recalled that date column includes both date and time of purchase, and then, after converting data as date type, splitting it in its corresponding columns - date and time - and deleting the former one, we decided to create also further attributes, which represents year, month, day, hour, week and weekday, to carry out more in-depth exploratory analysis.
Finally, as final step, we dropped the milliseconds from the newborn time column, to achieve a final format hh:mm:ss.

Then we continued preprocessing our data, in particular focusing ourselves on the deal of handling inconsistent numerical values.
In fact, taking a closer look at the dataset, we found out that we have negative values on the field that represents quantities, and zero values on the one showing sales prices.
We decided to deal with them in two different ways:

- As for negative quantities, which occur in only 9 instances, we simply decided to drop the corresponding rows, assuming quantities that assume values below 0 as clearly inconsistent data.

- Regarding the prices that are set equal to zero, in this case we have that the number of rows in which we encounter this issue is 1825, a number that, nevertheless, is significant in the context of our dataset, representing approximately 0.4% of the data.
  Therefore, needing to proceed with more caution, we perform a more detailed analysis on these data, as follows:

marchio	descrizione	settore	
Happy Runner Club	Canotta	Running	730
	Short	Running	493
	Vario	Running	493
Pdx Sport	Calze sportive	Calcio	56
Happy Runner Club	Varie abbigliamento	Running	48
Adidas	Vario	Calcio	1
Bodyline	Pallone	Volley	1
Dc Shoes	Scarpa	Snowboard	1
Happy Runner Club	T.shirt m/m	Running	1
Shimano	Mulinelli	Pesca	1

Figure 3. *Number of tuples with price set to 0, grouped by our categorical variables*

  For the null values with unit frequency, we attributed its cause to a tabulation error. In the case of a higher count of the missing values, we realized that it relate to the same brand. Since after a brief online research, we discovered that this brand is a non-profit association, we assume that the missing price could be due to some voluntary initiative or could be attributable to a commercial campaign.
  However, for all the missing values, we decided to exclude them from our dataset. For the unit null values, we opted for this choice because we believe that their frequency - they occur just 5 times - is small and not representative of the data set; for the second one, instead, although the occurrence is significant, we opted to proceed with the same procedure, because our goal is to analyze real sales with and effective currency exchange.

Then, after having completed our analysis on negative or null data, we focus on setting new attributes. First of all, we created a column named “Fatturato”, that we are willing to use through our work, in which we computed the multiplication between ’qta’ and ’prezzo’ columns.

Finally, we decided to add a further categorical attribute, called “categoria”, in we which organize the values of the column “settore” in more aggregated data.

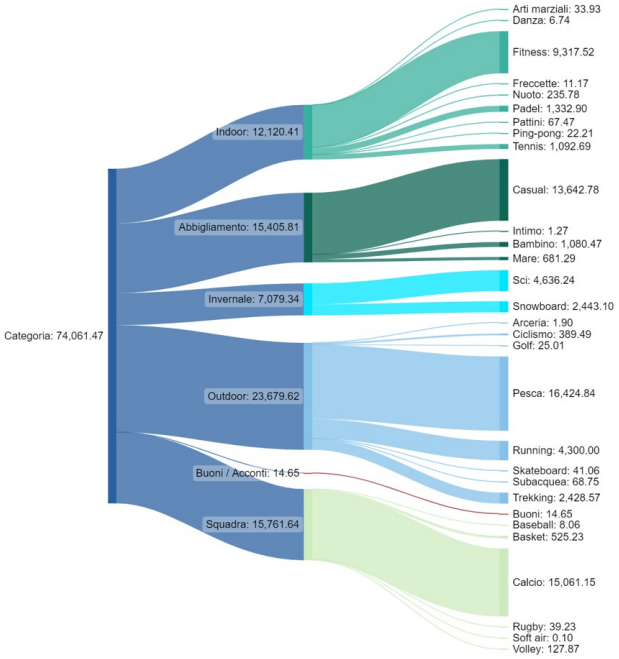

Figure 4. *A sankey diagram showing all the categories with their corresponding sectors*

Thus, in the sankey diagram, we have labelled each categoria with a distinctive color and the magnitude of the flow represents the total sale of sector over the reference period. To provide a better visualization, we changed the unit scale of sales to thousands.
According to the graph, it’s clear that, as for our store, the most profitable category is recorded to be the Outdoor one, which is worth approximately 23 700 thousand euros. Moreover, the Pesca sector was the best performing one in terms of sales (16 425 thousand euros), followed by Calcio (15 060 thousand euros) and Casual (13 643 thousand euros), which belong to “Squadra” and “Abbigliamento” categories, respectively.

2. Data integration

After completing the preprocessing phase, we proceeded by integrating the dataset with data retrieved from other sources, which can add information to our data and consequentially enrich the results we will develop with our predictive model.
More in detail, we wanted to include in our analysis factors that in our opinion could have conditioned both quantities of sold products and their revenues, through the years.
So we essentially decided to add to the preprocessed dataset a column for each of these fields.
Thus, in each of the upcoming sections we will present those factors and illustrate the corresponding covariate we added to our dataset.

2.1 Covid-19

The first element that came to our mind as a possible influencing factor for the store's sales trend is the Covid-19 pandemic, with its resulting restrictions.
Indeed, due to the COVID-19 pandemic, two phenomena occurred: firstly, the possibility of visiting a physical store was not always guaranteed (either due to travel restrictions or the closure of non-essential shops, such as sports stores); secondly, certain sports, especially team sports or outdoor activities, faced prohibitions or restrictions, leading people to shift out of necessity their focus to indoor activities that could be practiced within their homes.
Thus we wanted to check how and how much Covid-19 influence affected the sales trend of our e-commerce store, both in absolute terms and in more detailed terms, for example by distinguishing between the categories of sports created during the preprocessing phase, and we clearly focused ourselves exclusively in the last period reported in our dataset, that is from 2020 onward.
To do so, we retrieved a dataset in CSV format from Istituto Superiore della Sanità website [2], containing the number of hospitalizations for each day of the year, starting from 20th February 2020, the day before the identification of the first patient zero in Italy, i.e., the first individual in whom the presence of the virus was detected [3]. After dropping the last column, that just showed the date of the last update, the same for each row, and the last row, which just reported the total sum of all the hospitalizations through the years, we obtained a dataset with these features.

Dataset statistics

Number of variables	2
Number of observations	1183
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	18.6 KiB
Average record size in memory	16.1 B

Figure 5. *Overview of the Covid-19 hospitalizations dataset*

where the two variables are simply date and the corresponding number of hospitalizations. After merging the two data set on the date column, we renamed the variable of interest in the column "N_ ricoveri" to better capture this new information.
In the following figure, we compared daily Hospitalizations and total sales recorded by e-commerce. Most of the observations seem to be concentrated in the range between 50,000 and 15,000 thousand euros. The points are randomly arranged in the plane, showing an undefined pattern.

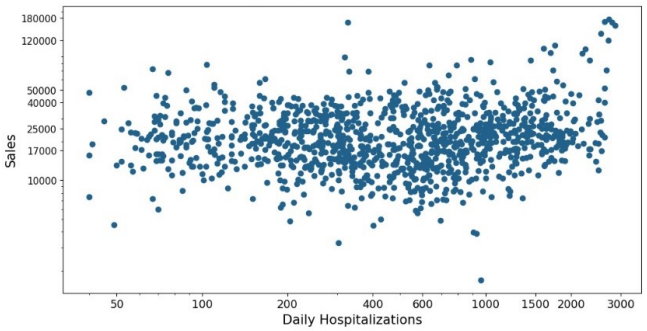

Figure 6. *Scatterplot that compares amount of sales with number of hospitalizations each day*

This suggests the presence of a slight association between the two variables. The relatively low correlation between the two variables, computed to be 0.26, supports this observation, indicating a positive but not strong association between daily hospitalizations and sales.

The next graph shows the sales fluctuation over our frame time. During the early stages of the pandemic emergency, sales, especially those in the fitness sector, experienced an unpredictable surge.

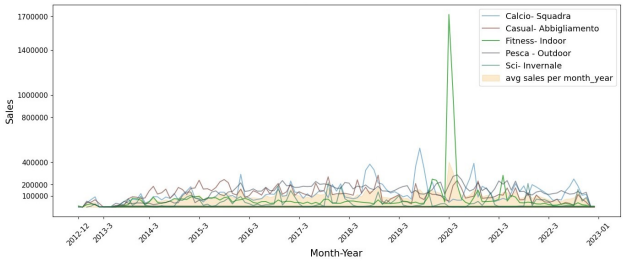

Figure 7. *Sales Distribution by Sector (2022-02 to 2023-04)*

To provide context, this refers to the period between March and April 2023, during the first phase of the lockdown. With the decree issued on 9 March, all travelling outside one's municipality was banned across the country. Additionally, on 11 March, a new DPCM (Decree of the President of the Council of Ministers) was published, which provided for the closure of all activities not deemed necessary, such as recreational economic activities [4]. As a result, there was a notable increase in the demand for personal training products and accessories, such as home fitness equipment.

2.2 Winter and summer markdowns and Black Fridays

Another factor that we consider as a potential influencing driver on sales is periodic markdown events. These periods represent times of the year when companies liquidate their remaining stock at significant discounts.
In our variables we decided to include Black Friday as well as winter and summer markdowns. The retrieval of this information was a challenging task, which we did not expect. Whether

the definition of the day of the Black Friday is clear, occurring on the Friday after Thanksgiving in the United States; the markdowns dates vary from year to year and are subject to decisions by local or regional authorities.
To establish a comprehensive and nationally applicable period, we have considered the most representative dates provided by Confcommercio [5] and Confesercenti [6]. Typically, both last approximately 60 days: winter markdowns tend to start in early January and end in early March, while summer markdowns start in early July and end in early September. It must be said that 2020 was an exceptional year, since due to the pandemic emergency, the summer markdowns was postponed from its usual start date on 1 August until the end of September. To capture this information effectively, a binary variable called "Saldi" was then created, which assigns a value of 1 to indicate a markdowns day and zero otherwise.
Our assumption about Black Friday markdowns have been fairly confirmed: 6 out of 10 days with the highest amount of sales for November effectively correspond with the day we expected. This trend is true for the years 2016-2021.

Moving on, the following graph illustrates the total sales per week of the year. Notably, there is a peak before Christmas, during the 50th week, when e-commerce records its highest sales.
Regarding instead the initial weeks of the year relating to the winter markdowns, there is evidence of a robust sales surge only in the first stage, while the tendency gradually decreases as the days go by.

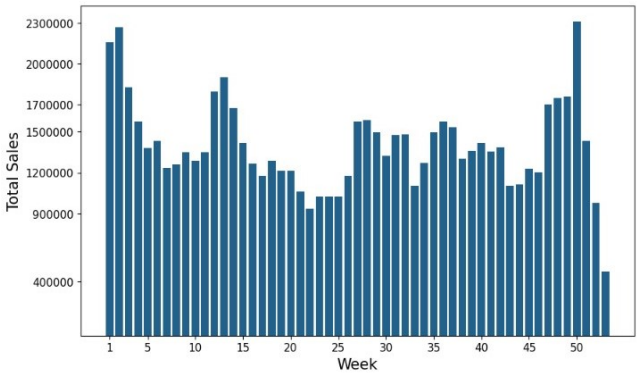

Figure 8. *Total Sales per Week*

In contrast, the summer markdowns, which typically occur between the 27th to 36th week, exhibit a u-shape curve: with sales that reach 1,5 Million in July, just before the summer break, with this threshold being exceeded again in the first week of September, probably corresponding to the return to everyday life and the resumption of sports facilities and activities.

2.3 Holidays

We believed that holidays could be a influencing factor that could have positively impact on sales. To determine which day should be labelled as festivity, we employ the "holidays" library [7], that considers a day as holiday whether it is a non-working day. Before further proceeding, we have to report that some festivities have fixed dates, while others, such as Easter and Easter Monday dates, vary each year. Therefore, a Boolean variable called 'Holiday' was created to effectively capture this information, assigning the value 1 to indicate a holiday and zero otherwise. On the following snapshot, we show the holidays referring to the year 2022.

```
2022-01-01 Capodanno
2022-01-06 Epifania del Signore
2022-04-17 Pasqua di Resurrezione
2022-04-18 Lunedì dell'Angelo
2022-04-25 Festa della Liberazione
2022-05-01 Festa dei Lavoratori
2022-06-02 Festa della Repubblica
2022-08-15 Assunzione della Vergine
2022-11-01 Tutti i Santi
2022-12-08 Immacolata Concezione
2022-12-25 Natale
2022-12-26 Santo Stefano
```

Figure 9. *The list of the holidays we consider in 2022*

To test our hypothesis, we computed the average sales over the period from 2012 to 2023. According to our initial thought, festivity days should register an higher amount of orders, do to the fact that they coincide to the days in which physical stores tend to be closed.
On the contrary, surprisingly we found out that the customers tend to place more orders on non-holiday days compared to on holidays. Specifically, the average daily sales on non-holiday days amounted to 20 331 thousand euros, while the average daily sales on holidays were just 17 631 thousand euros. These results indicate that holidays might have a negative impact on sales, on the contrary to what we originally supposed.

3. Data visualization

In this section, we present additional graphical insights that we have not had the opportunity to introduce in the previous sections and which may be useful to better understand the potential of the overall dataset, to identify prospective business opportunities for products or brands, and finally to make informed medium- and long-term decisions to consciously drive the revenue growth.

Firstly, we created a stacked bar chart representing the share of sales by sector displayed for each year, excluding the partial years of 2012 and 2023 from the dataset. This allows to draw conclusions while avoiding erroneous assumptions.
However, we can observe that the two dominant sectors, Pesca and Calcio, consistently drive most sales. They account together for approximately 30% of share each year: a profitable combination that can tell a lot about interest and demand of the average customer.
On the other hand, the casual sector appeal seems to have

declined over the years. After peaking in 2018 with a market share of 28%, the sector's contribution has steadily declined to below 10% in 2022.

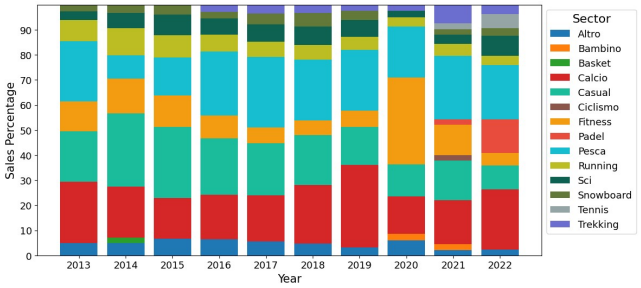

Figure 10. *Percentage of sales for each sector during the years of analysis*

As already reported, 2020 was an unusual year: the e-commerce recorded an year-on-year increase of Fitness sector of 708% moving from 487 K euros of sales to even 3.9 M euros; no other sector sees an increase of this magnitude, but, as we can see looking at the following years, this exceptional behavior has not turned into a trend.
Two other sectors that are worthy to be described are Tennis and Padel, which accounted respectively for a share of 5.5% and 13.9% in 2022: they generated an additional increase of 718 and 178 thousands of euros in sales, compared to the previous year. This suggests that these sectors have the potential for further growth and should be monitored closely.

In this second graph, we present a heatmap that illustrates the distribution of sales based on day and hour of the day.
Overall we can observe that the majority of purchases are made during daylight hours, with most concentration of sales between the time zones 11-12, 14-16 and 21-22 on workdays. Conversely, during weekends, customers tend to prefer making purchases after 17.

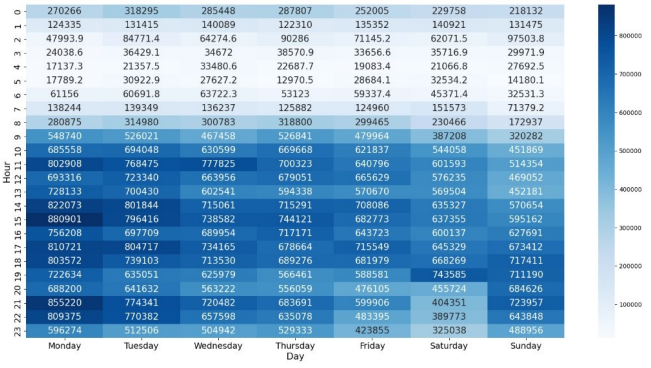

Figure 11. *Sales Heatmap on Hours and Days of the Week*

When considering specific days of the week, an interesting trend emerges. It seems that Monday is not just the day of returning to work but coincides also with the actual purchase of the products. From a profitability point of view, Monday emerges as the most lucrative day. This behaviour persist until Tuesday, then fades away throughout the week and resumes its cycle the following Monday. This graph highlights the importance of targeting specific time intervals for advertising campaigns and capitalizing on these purchasing trends.

In this third graph we show the total sales per brand and the respective market share in e-commerce. It is noteworthy that Adidas and Nike are the only sectors that generate more than 5 million euros in sales, respectively with 10.9% and 10.12% of the total turnover. These brands offer a wide range of products related to various sports, but in our e-commerce, their offerings extend to indoor and outdoor categories, as well 'abbigliamento' and 'squadra' items.

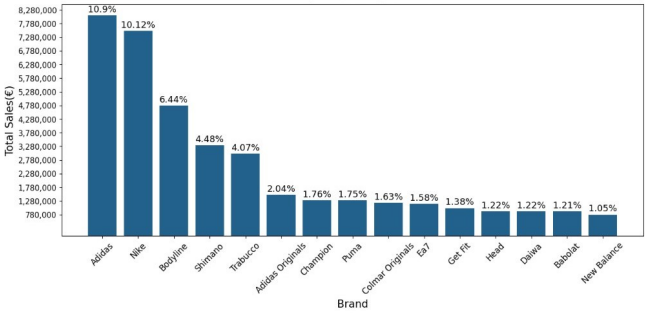

Figure 12. *Best performing Brand*

Moving on, on the lowest step of the podium we can find the brand Bodyline, which accounts 4.8 million euros and offers products belonging to all our five macro-categories: in addition to those mentioned above, even the winter one.
By focusing on single-sector brands, Shimano, Trabucco and Daiwa emerge. These brands solely focus on the Pesca sector and collectively achieve a 9.77% market share.
Another interesting brand is Colmar Originals, which besides offering casual, children's clothing and beachwear, is also specialized in winter sports clothing such as wind jackets, fleece coats in pile and padded trousers.

Finally, the last graph reveals which is the product that has been sold the most during our referring period. To add further information, each product is colour-coded according to the sector it belongs to.

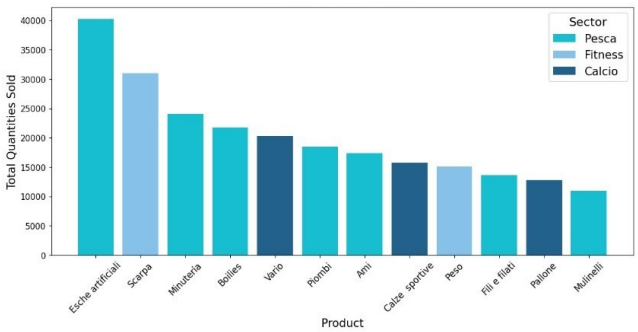

Figure 13. *Most Sold product*

Analyzing the graph, we can state that he best-selling products belongs to just three sectors: Calcio, Pesca and Fitness. Among these, Pesca leads with seven out of twelve articles represented.
Changing our perspective, the top performer product results to be 'Esche Artificiali', which recorded sales that amount to more than 40 thousands units, followed by 'Scarpe', with approximately 30 thousands articles, and 'Miniteria', with 25 thousands articles.
Being aware that these current sectors play a key role in driving the sales, the management probably has to think about strategies that involve their prioritization, in order to keep the business model sustainable and profitable in the long-term.

4. Models

The goal of this project is to use the dataset at our disposal as a time series of e-commerce turnover to train forecasting models. The models we chose to predict future revenue were trained and used to predict both daily and weekly revenues.
First of all, therefore, two different datasets were created. We list them below:

- Ecommerce_giornaliero: this dataset was created by grouping the original dataset by day, month and year. The 'Fatturato' column was added while the 'N_ricoveri', 'Saldi' and 'Festa' columns, obtained after the data integration phase, remained unchanged
- Ecommerce_settimanale: this dataset was created by grouping the ecommerce_giornaliero dataset by year and month. The 'Fatturato' column has been added together, as well as 'N_ricoveri', 'Saldi' and 'Festa' columns.

Before proceeding with the implementation of the predictive models, we verified through the Dickey-Fuller [8] and KPSS [9] tests that the series was stationary in mean and in variance.

4.1 ARIMAX

The ARIMAX model [10], acronym for Autoregressive Integrated Moving Average with Exogenous Variables, is an extension of the ARIMA model that allows you to incorporate exogenous variables into the forecasting process. In general, the ARIMAX model is expressed in the form:

$$\begin{aligned} Y(t) &= c + \phi_1 \cdot Y(t-1) + \ldots + \phi_p \cdot Y(t-p) + \\ &\quad + \theta_1 \cdot \varepsilon(t-1) + \ldots + \theta_q \cdot \varepsilon(t-q) + \\ &\quad + \beta_1 \cdot X_1(t) + \ldots + \beta_k \cdot X_k(t) + \varepsilon(t) \end{aligned}$$

where:

- $Y(t)$ represents the dependent variable or the time series to be forecast.
- c is a constant term.
- $\phi_1, \ldots, \phi_p$ are the coefficients of the auto-regressive terms.
- $\theta_1, \ldots, \theta_q$ are the coefficients of the moving average terms.
- $\varepsilon(t)$ is the white noise or residual error.
- $X_1(t), \ldots, X_k(t)$ are the exogenous variables or covariates to be included in the model.
- $\beta_1, \ldots, \beta_k$ are the coefficients of the exogenous variables.

The main objective of the ARIMAX model is to capture the linear relationships between the dependent variable and the exogenous variables, in order to improve the accuracy of the predictions. Exogenous variables may represent external factors or influences which may affect the historical series and which may not be incorporated into the auto-regressive or moving average pattern.
In our case, the exogenous variables that can have an influence on the model are:

- 'Festa': boolean variable that indicates whether the day in question is an Italian public holiday or not.
- 'N_ricoveri': integer type variable indicating the number of patients hospitalized by Covid on that day
- 'Saldi': Boolean variable that indicates whether there are balances on the day in question.

To determine the parameters of the ARIMA model, the RMSE comparison method was used by varying the parameters ϕ and θ from 0 to 5 and setting d always equal to 0 since no differentiation was made.
The best result according to the RMSE criterion is the following pattern (4,0,3), this means that:

- p = 4: indicates that the ARIMA model includes two auto-regressive terms, i.e. it relies on the two previous values of the time series to predict the current value.
- d = 0: indicates that no differentiation is applied to the historical series. The series is considered stationary or has previously been made stationary.
- q = 3: indicates that the ARIMA model includes a moving average term, i.e. it relies on the previous residual value to improve the prediction.

We will now report the coefficient results of the trained ARIMAX model:

- const: 17965.50
- N_ricoveri: 1.740011
- saldi: 2632.366
- festa: - 917.4142

- ar.L1: 0.4060158
- ar.L2: - 0.08053035
- ar.L3: 0.9827067
- ar.L4: - 0.3176978
- ma.L1: 0.1457872
- ma.L2: 0.2427603
- ma.L3: - 0.7738534
- sigma_2: 6.807324e+07

So the final equation is:

$$\begin{aligned} Y(t) \;=\; & 17965.50 + 1.740011 \cdot N_{\text{ricoveri}}(t) + 2632.366 \cdot \\ & \cdot \text{saldi}(t) - 917.4142 \cdot \text{festa}(t) + 0.4060158 \cdot \\ & \cdot Y(t-1) - 0.08053035 \cdot Y(t-2) + 0.9827067 \cdot \\ & \cdot Y(t-3) - 0.3176978 \cdot Y(t-4) + 0.1457872 \cdot \\ & \cdot \varepsilon(t-1) + 0.2427603 \cdot \varepsilon(t-2) - 0.7738534 \cdot \\ & \cdot \varepsilon(t-3) + \varepsilon(t) \end{aligned}$$

where:

- $Y(t)$ represents the dependent variable or the time series to be forecasted
- '$N_{\text{ricoveri}}(t)$' represents the exogenous variable 'N_ricoveri' at time t
- 'saldi(t)' represents the exogenous variable 'saldi' at time t
- 'festa(t)' represents the exogenous variable 'festa' at time t
- $\varepsilon(t)$ is the white noise or residual error at time t.

The coefficients const, 'N_{ricoveri}', 'saldi', 'festa', ar.L1, ar.L2, ar.L3, ar.L4, ma.L1, ma.L2, ma.L3 correspond to the estimated values from the ARIMAX model.
Therefore, according to the model, the company's turnover decreases by 917 euros during holidays and also decreases by

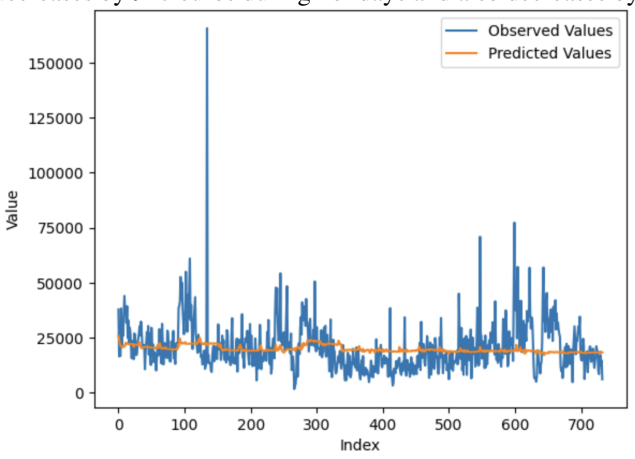

Figure 14. *Comparison between expected values and observed values of the ARIMAX model with daily time series*

The result is that the model tends to underestimate the observations as can be seen in the figure. In general, the model has a fairly constant decreasing trend over time, which foresees peaks on sales days. The turnover trend for the month from 2023-04-04 to 2023-05-04 predicted by the model will be shown, taking into consideration the four holidays present in this range of dates.

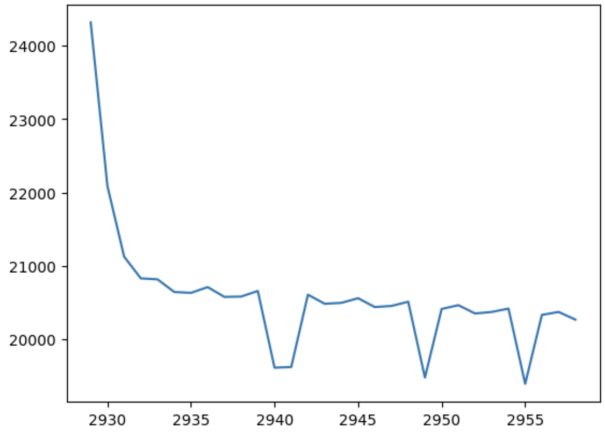

Figure 15. *Forecast for next month ARIMAX model*

Then we analyzed the residuals:

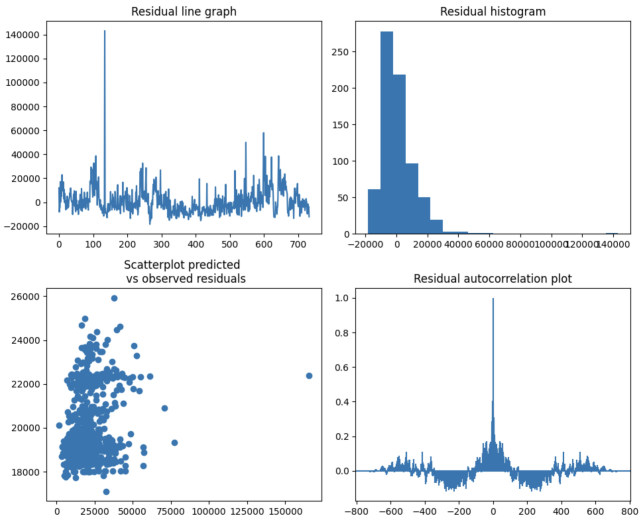

Figure 16. *Residue analysis of ARIMAX model*

From the analysis of the residuals it emerges that the data could be affected by heteroskedasticity since the histogram presents a strong asymmetry. Of course, the nature of these anomalous peaks is due to the increase in daily e-commerce turnover in times of Covid-19. In fact, during the pandemic, the volatility of the data has increased considerably and consequentially the variance of the errors depends on the period considered and on the presence of anomalous events.

4.2 Prophet

The Prophet model [11] is a forecasting model developed by Facebook for forecasting time series. It is widely used for forecasting temporal data, especially data that has seasonal trends and anomalies. The Prophet model is based on an additive approach in which the historical series is decomposed into trend, seasonality and error components. This model offers many features that simplify the forecasting process, such as automatic trend management, holiday detection, and the flexibility to add exogenous components.

Using the Prophet model to forecast the daily turnover of e-commerce we obtained the following results in terms of series decomposition:

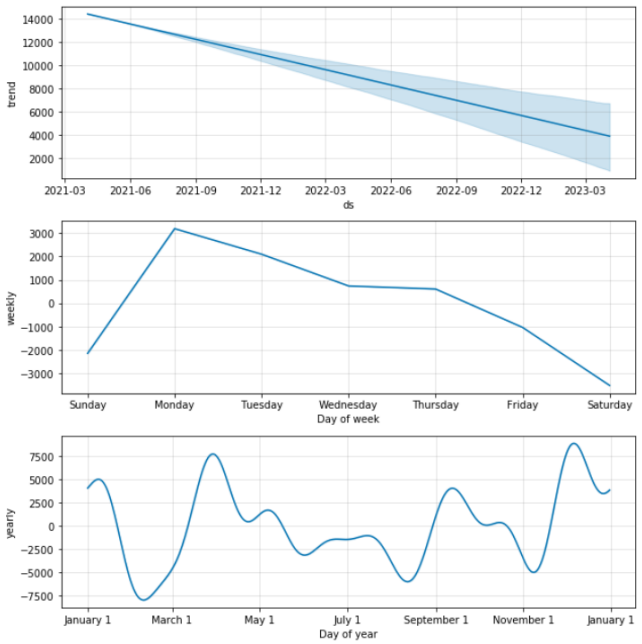

Figure 17. *Analysis of the components of the time series*

As we can see the breakdown of the historical turnover series, it indicates a downward trend probably due to the reopening of physical shops post corona. Furthermore, a seasonality is highlighted in the months of September, December and January; the increase in sales in recent months could be due to the arrival of Christmas and the resumption of sport after the summer holidays for both adults and children. Finally, from the decomposition, it emerges that Monday seems to be the day when people make the most purchases.
Therefore, the result we expect from the predictions of the Prophet model is an increase in turnover on the days corresponding to Monday and in the months of September, December and January. Furthermore, in general we expect a downward trend.

Exactly as in the previous point, also this time we used the covariates 'N_ricoveri', 'Saldi' and 'Festa', and the dataset was divided with a proportion of 80% for the training set and 20% for the test set with the following results:

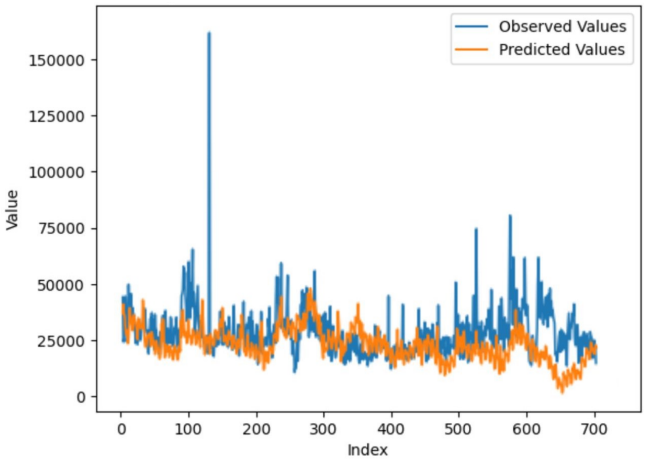

Figure 18. *Comparison between expected values and observed values of the Prophet model with daily time series*

As you can see the Prophet model seems to fit the data much better than the ARIMAX. The Prophet model handles seasonality better than the ARIMAX model, in fact the Prophet offers a robust implementation of seasonality modeling, allowing you to accurately model and predict seasonal patterns over time series.

The residuals analysis of the Prophet model will be shown below:

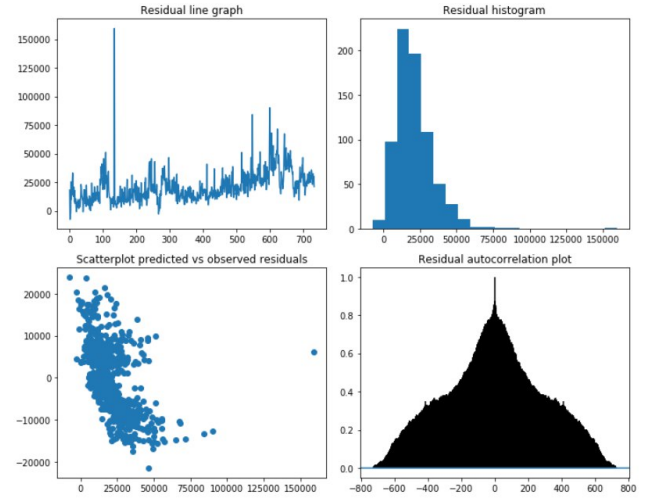

Figure 19. *Analysis of the residuals*

Also this time, the analysis of the residuals shows us the possible presence of heteroskedasticity among the data, clearly highlighted in the graph of expected residuals vs observed residuals. The grouping of all the observations towards the left side of the Cartesian plane could also indicate that the model is not correctly considering some relevant factors or variables that influence the time series.

We will now show the predictions of the Prophet model for the next 30 days

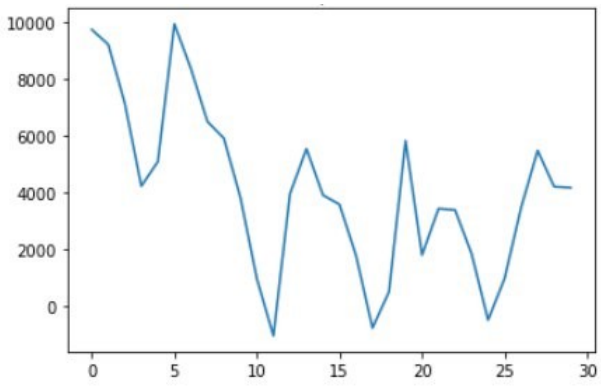

Figure 20. *Forecast for next month Prophet model*

As can be seen, the positive peaks are on Mondays while the trend is decreasing for the rest of the days. The negative peaks, on the other hand, correspond to holidays, since the model predicts a drop in turnover for holidays.

4.3 XgBoost

XGBoost, short for "eXtreme Gradient Boosting" [12], is a machine learning algorithm based on the concept of boosting, which combines several weak models to create a complex and powerful model. It is widely used for regression and classification problems.

The XGBoost model is based on the gradient boosting technique, which is an ensemble learning algorithm. The goal of gradient boosting is to build a predictive model by combining a set of weak models, called decision trees, sequentially. Each weak model is trained to correct the residual errors of the previous model.
XGBoost distinguishes itself by using an optimization algorithm that improves efficiency and accuracy of gradient boosting. This algorithm is called gradient boosting with the use of approximate gradients and an efficient implementation of the best-splits-finding algorithm to build the trees.

We trained with our data an XgBoost model with 'N_ricoveri', 'Saldi' and 'Festa' as covariates, and with these input parameters:

- *Max_depth*: The maximum depth of the decision trees. It can be used to control the complexity of the model and prevent overfitting. We set this parameter to 3 to help prevent overfitting and reduce model complexity, and also because we have 3 covariates.
- *Learning_rate*: The learning rate of the model. It controls the importance given to each tree in updating the weights. We set this parameter to 0.1 to slow down model fitting, giving more weight to regularization and reducing the risk of overfitting.
- *N_estimators*: The number of decision trees to create in the model. We set this parameter to 25 to balance model complexity with training time.
- *Reg_alpha & Reg_lambda*: The L_1 and L_2 regularization terms, respectively. They can be used to apply a penalty to the model's weights and control the complexity. We set these parameters to 0.01 to encourage the model to prefer simpler solutions and smaller coefficients and to reduce the risk of overfitting.

The result is showed below:

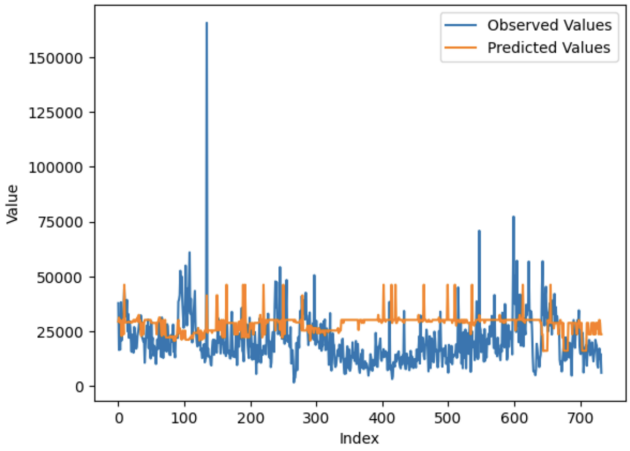

Figure 21. *Comparison between expected values and observed values of the XgBoost model with daily time series*

Looking at the graph, it can be seen that the XgBoost model has a fairly constant trend over time with data overestimations.

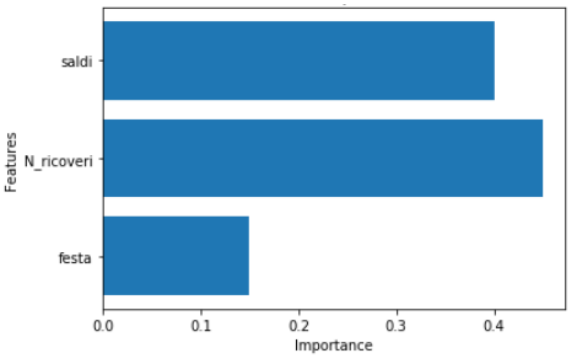

Figure 22. *Importance of covariates*

From the analysis of the importance of the covariates and from a more in-depth analysis of the 25 decision trees, it emerged that the trained XgBoost model gives much more importance and relevance to the number of hospitalized by coronavirus. On the other hand, it gives less importance to sales and holidays. In fact, we would have slight fluctuations in turnover in correspondence with sales and holidays, while we would have

larger fluctuations as the number of hospitalized patients from the coronavirus increases.

Now let's move on to the analysis of the residuals:

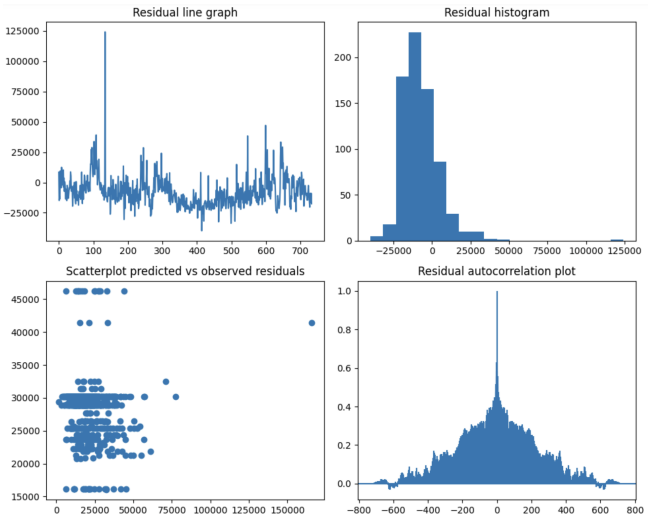

Figure 23. *Analysis of the residuals*

The result of the residual analysis is very similar to that seen for the previous models, so we will not discuss it further.

4.4 Multivariate Linear Regression

Multivariate linear regression is a statistical technique used to model the relationship between a continuous dependent variable and two or more independent variables [13]. In other words, it is a question of extending the concept of simple linear regression, in which only one independent variable is considered, to a context in which there are multiple independent variables.
The equation of a multivariate linear model can be represented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Where:

- Y represents the dependent variable.
- $X_1, X_2, \ldots, X_n$ represent the independent variables.
- β_0 represents the intercept term or the constant.
- $\beta_1, \beta_2, \ldots, \beta_n$ represent the coefficients or weights associated with the independent variables.
- ε represents the error term or residual.

This equation expresses the linear relationship between the dependent variable and the independent ones, where the coefficients (β_i) represent the overall change in the dependent variable for a unit change in the corresponding independent variable, holding all other variables constant.

In our case, the dependent variable is the turnover while the independent variables are 'N_ricoveri', 'Saldi' and 'Festa'. The estimated regression coefficients are as follows:

- Intercept: 17 965,500 798 025 514
- N_ricoveri: 11,319 467 51
- Saldi: 2 632,364 609 4
- Festa: −917,413 708 89

The multivariate linear regression model therefore predicts a growth in daily turnover of 2 632,36 euros on sales days while it predicts a loss of 917,41 euros on holidays. Finally, the model provides for an increase in turnover of 11,31 euros for each hospitalized by the coronavirus.

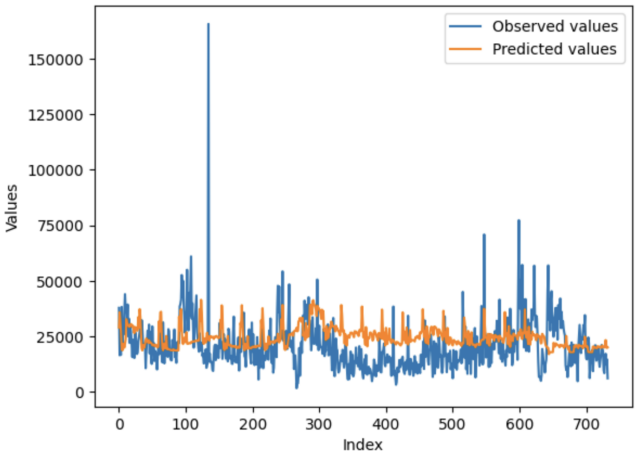

Figure 24. *Comparison between expected values and observed values of the Linear model with daily time series*

So the model equation is:

$$\begin{aligned} Y \;=\; & 17965.500798025514 + \\ & + 11.31946751 \cdot \text{N\_ricoveri} + \\ & + 2632.3646094 \cdot \text{Saldi} + \\ & - 917.41370889 \cdot \text{Festa}. \end{aligned}$$

Then, also in this latter case, we proceeded with the analysis of the residuals, on which we will not dwell, as they present similar situations to the previous cases; thus we limit ourselves to show the graphical results:

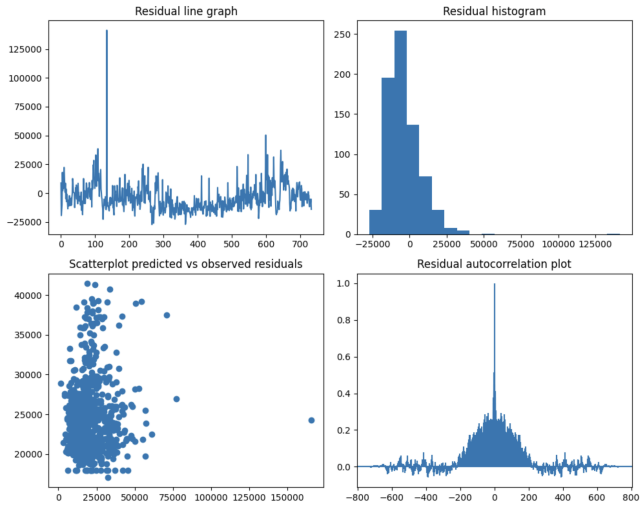

Figure 25. *Residual analysis of XgBoost model*

4.5 SARIMAX

The SARIMAX model, which stands for Seasonal Autoregressive Integrated Moving Average with eXogenous variables, is a time series forecasting model that incorporates both seasonal and exogenous components [14]. It is an extension of the ARIMA model that allows for the inclusion of additional variables that may influence the time series.

The SARIMAX model takes into account the seasonal patterns in the data, which are characterized by repeating patterns at fixed intervals. It captures the auto-regressive (AR) component, which represents the relationship between the current observation and previous observations, the moving average (MA) component, which captures the relationship between the current observation and the residual errors from previous observations, and the integrated (I) component, which accounts for trends and stationarity in the data.

In addition to the seasonal and auto-regressive components, SARIMAX models can include exogenous variables, which are external factors that may impact the time series. These variables can be included to improve the accuracy of the forecast by considering their influence on the dependent variable.

In our model we used the usual covariates 'N_ricoveri', 'Saldi' and 'Festa' and as parameters we used the same as the ARIMAX model, therefore p=3 and q=4. We set the seasonal order componment in this way:

- The seasonal autoregressive (SAR) component has an order of 1. It captures the relationship between the current observation and the previous observation at the same seasonal lag (lag 30).
- The seasonal differencing (S) component has an order of 1. It indicates that the series needs to be differenced once at the seasonal lag (lag 30) to achieve stationarity.
- The seasonal moving average (SMA) component has an order of 0. It implies that the model does not include a seasonal moving average component.
- The seasonal cycle length is 30, suggesting that the seasonal pattern repeats every 30 observations, so every month.

After training the model, the equation is given by:

$$\begin{aligned} Y(t) \;=\; & 2.510380 \cdot \mathrm{N}_{\mathrm{ricoveri}}\,(t) + 2.528636 \cdot \text{ saldi }(t) + \\ & - 2.588355 \cdot \text{ festa }(t) - 0.07886526 \cdot Y(t-1) + \\ & + 0.9232968 \cdot Y(t-2) + 0.03969555 \cdot Y(t-3) + \\ & + 0.635876 \cdot \varepsilon(t-1) - 0.532307 \cdot \varepsilon(t-2) + \\ & - 0.2378532 \cdot \varepsilon(t-3) - 0.06114475\varepsilon(t-4) + \\ & - 0.4916517 \cdot Y(t-30) + \varepsilon(t) \end{aligned}$$

where:

- $Y(t)$ represents the dependent variable or the time series being forecast.
- '$\mathrm{N}_{\mathrm{ricoveri}}\,(t)$' represents the exogenous variable "N_ricoveri" at time t.
- 'saldi'(t) represents the exogenous variable "saldi" at time t.
- 'festa'(t) represents the exogenous variable "festa" at time t.
- $\varepsilon(t)$ is the white noise or residual error at time t.

The coefficients saldi, festa, $\mathrm{N}_{\mathrm{ricoveri}}$, ar.L1, ar.L2, ar.L3, ma.L1, ma.L2, ma.L3, ma.L4, and ar.S.L30 correspond to the estimated values from the SARIMAX model. The result of the model prediction is shown in the following figure:

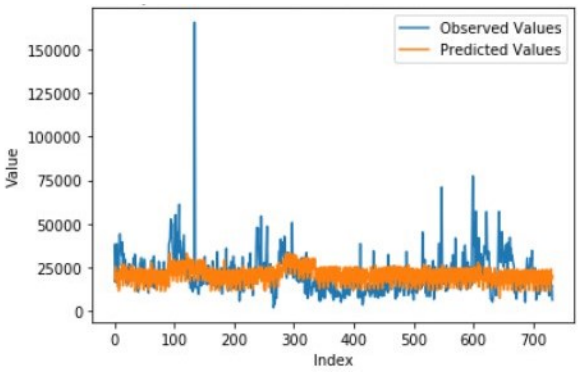

Figure 26. *Comparison between expected values and observed values of the Sarimax model with daily time series*

The SARIMAX model predicts data much better than other models because it places a lot of weight on monthly seasonality. This model also predicts data with a decreasing trend.
We now show the graphical representations of the residuals, once again without further comment:

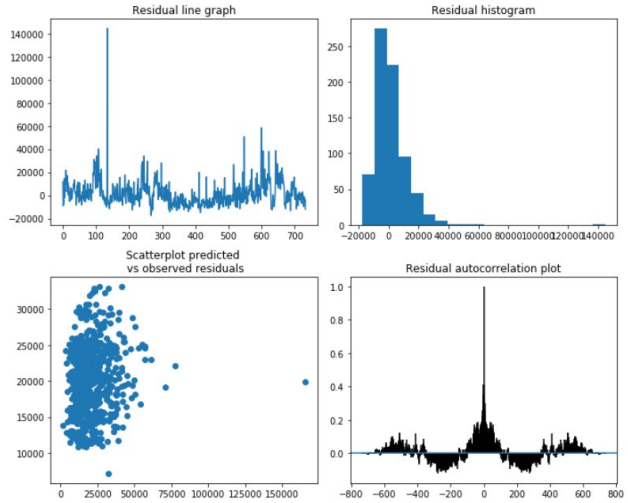

Figure 27. *Residual analysis of SARIMAX model*

5. Evaluations

Once the models were trained, various evaluation metrics were used to see which of these was the best one. We will present them in different dedicated sections, then we will create a further one in which we will present the best models according to these metrics.

5.1 Mean Absolute Error

The MAE calculates the mean difference between the predicted values and the observed values, imposing the sign of the differences as positive.
Mathematically, the MAE is calculated as the average of the sum of the absolute errors:

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

where y_i represents the observed values, $\hat{y}_i$ represents the expected value and n represents the total number of observations. The MAE measures the mean error in absolute units and provides an estimate of the mean deviation between the predicted and observed values. The lower the MAE value, the better the model performance.

5.2 Mean Squared Error

The MSE calculates the mean squared differences between the predicted values and the observed values. This metric penalizes larger errors more than smaller errors, due to the square in the formula.
Mathematically, the MSE is calculated as the mean of the sum of squared errors:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i represents the observed values, $\hat{y}_i$ represents the expected value and n represents the total number of observations. The MSE provides a measure of the mean squared error between the predicted and observed values. A lower value of the MSE indicates a better fit of the model to the data.

5.3 Root Mean Squared Error

The RMSE provides an estimate of the mean error between the predicted and observed values. Because it is calculated as the square root of the MSE, the RMSE has the same unit of measure as the original data, which can make it easier to interpret.
Mathematically, the RMSE is calculated as the square root of the MSE:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

Like the MSE, a lower value of the RMSE indicates a better fit of the model to the data. The RMSE is especially useful when you want to compare the mean error across different models, or when you want to interpret the error in the context of the units of the original data.

5.4 Mean Absolute Percentage Error

MAPE is calculated as the average of the absolute values of the percentage differences between the predicted and observed values, expressed as a percentage:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

where y_i represents the observed values, $\hat{y}_i$ represents the expected value and n represents the total number of observations. The MAPE provides a measure of the mean percent error, allowing you to evaluate how much the error of the predicted values deviates from the observed values on average. A lower MAPE value indicates a better fit of the model to the data.

It is important to note that MAPE can become problematic when there are observed values very close to or equal to zero, as it can cause divisions by zero. Also, the interpretation of percent error can vary depending on the context and units of measurement of the data.

5.5 Best Models

Now, the evaluation metrics associated with each of the four models described above will be reported.

Modello	MAE	MSE	RMSE	MAPE
ARIMAX	7570.03	1.246289e+08	11163.73	49.58
Prophet	21062.97	6.215704e+08	24931.31	101.20
XgBoost	11846.25	2.090145e+08	14457.33	88.32
LM	9706.01	1.594794e+08	12628.52	66.92
SARIMAX	8341.57	1.420094e+08	11916.77	50.88

Figure 28. *Final comparison of the values of all the considered metrics for each model taken into account*

Looking at model evaluation metrics, ARIMAX and SARIMAX models are the ones that fit the data best. The reason

is due to the nature of our dataset which contains a historical series of 10 years and therefore the best predictive models are those for historical series that take into account the delays and past correlations.
The Prophet, XgBoost, and linear regression models don't fit the data very well since they rely heavily on the influence of covariates. The covariates available in our dataset were not completely exhaustive since the purchase of sports products can be influenced by many factors.

Conclusions and future developments

The goal of this project was to find the best predictive model in order to predict as precisely as possible the future sales of a shop, in this case an online sales shop that mainly focuses on sports articles. Ad-hoc covariates that could positively influence the turnover have been added and as it has been seen, the covariate that most influences the turnover of an e-commerce resulted to be the one concerning the number of hospitalizations, immediately followed by the covariate linked to the annual balances.

During the project it was tried several times to obtain data regarding the average Italian temperature, to add as covariate to be a basis for exploring trends and correlations between the external temperature and the number and type of purchases. We tried to search the data we were looking for through many different sources, but we could not find complete data regarding the whole nation and that could cover the entire time span. Since we had no information regarding the location of our e-commerce headquarters or the region or area in Italy from which the highest number of orders is placed, we could not consider temperatures which only refers to particular cities or regions, which was the best we could do, at least without having to pay to access the data.
In the end, for all this reasons we decided not to add to the dataset that new covariate regarding temperature.

Anyway, the results obtained by our models can be considered satisfactory, especially for historical series models such as ARIMAX and SARIMAX.

Some further developments could consist in adding more covariates to our dataset and consequently to our models, and in trying to train them again. One covariate to consider could be, in addition to the previously mentioned average temperature in Italy, the presence or absence of sports events of particular media relevance, such as the Olympics or Football World Cup, which can influence and encourage consumers to engage in sports or specific types of sports.

References

[1] Pandas profiling report: *https://pypi.org/project/pandas-profiling/*

[2] Sorveglianza Covid-19: i principali dati nazionali (iss.it): *https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza-dati*

[3] *https://www.fondazioneveronesi.it/magazine/articoli/da-non-perdere/covid-19-la-pandemia-in-10-date-da-ricordare*

[4] *https://global-monitoring.com/gm/page/events/epidemic-0001933.mgRcH9vjlVrX.html?lang=en*

[5] *https://www.confcommercio.it/-/saldi*

[6] *https://www.confesercenti.it/*

[7] *https://pypi.org/project/holidays/*

[8] *https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/*

[9] *https://www.statisticshowto.com/kpss-test/*

[10] ARIMAX model: https://github.com/vighneshutamse/ARIMAX/blob/master/ARIMAX.ipynb

[11] Prophet: *https://cran.r-project.org/web/packages/prophet/prophet.pdf*

[12] Xboost: *https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/*

[13] Multivariate Linear Regression: *https://towardsdatascience.com/multivariate-linear-regression-in-python-step-by-step-128c2b127171*

[14] SARIMAX model: *https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html*