Investigation of Holidays effect on restaurants Sales in Northen Italy: A time series analysis

Ambrosini Eloisa, Chabib Salma, Grasso Lilia, Internò Christian, Maffi Claudio
Department of Data Science,
University of Milan-Bicocca
e.ambrosini2@campus.unimib.it, s.chabib@campus.unimib.it,
l.grasso8@campus.unimib.it, c.interno@campus.unimib.it
c.maffi4@campus.unimib.it

***Abstract*— A restaurant sales prediction allows the owners to know when it is better to open a new site, when to replenish stocks and to get a better idea of the restaurant revenues. Moreover, the analysis has been carried out in order to investigate the festivities at which the restaurants may have the highest profits to help control inventory and staff wisely, and to predict earnings. Sales forecasting also has the capacity to affect every major decision in a restaurant, concerning opening days, especially in the troubled periods as COVID-19. In order to perform an accurate investigation, we clustered the restaurants in two groups that have the same quantity of daily sales during time, with the purpose of avoiding redundancy. We have compared different models among them, from SARIMAX to TBATS, using also the most innovative techniques developed by Facebook, such as Prophet, taking advantage from Python libraries as pmdarima. Thanks to these tools we have found out unexpected developments, for instance the incredible revenues achieved on the 15th of August by some of the restaurants. Furthermore, the losses during the major holidays are surprisingly huge, and this could lead owners to contemplate the idea of holiday-closings. What are the other holidays that influence in a significant way the revenues of an inland-located restaurant?**

***Index Terms*— Prophet, Restaurants' sales, Forecasting, Covid19, Comparison, Holidays**

I. INTRODUCTION

A time series is a sequence of data points that occur in successive order over some period of time. It tracks the movement of the chosen data points, such as a restaurant's sale, over a specified period of time with data points recorded at regular intervals, for example a data per each day. There is no minimum or maximum amount of time that must be included, allowing the data to be gathered in a way that provides the information being sought by the investor or analyst examining the activity. In this paper, it is analyzed a time series of daily restaurant revenue over a period of 5 years circa, from 1st January 2017 to 12th of April 2021. The analysis is carried out using technical analysis tools in order to know, for example, whether the restaurants' time series shows any seasonality to determine if the restaurants go through peaks and troughs at regular times each year; or taking the observed profits and correlating them to a chosen holiday (using Python's of libraries that associate date to holidays such as Christmas and New Year's Eve to respectively 25th and 31st of December and other Italian national festivities). A restaurant sales forecast enables the owners to know when is the best time to open a new location, what inventory to order for the next month, how many employees should be scheduled for a shift in two weeks and to get a better idea of the restaurant potential profitability. With a backlog of data and a solid sales forecast, it would be possible to predict sales for certain scenarios – like holidays, seasons, events, hot or cold weather, etc. Reviewing data, a stretch of cold weather may be spot as the catalyst for an increase in business at a cozy restaurant. By referring to last winter's sales data, it will be possible to see how costs fluctuate when the holidays come, and managers will be prepared to capitalize on the opportunity with ample staffing and supply levels. In order to carry out a proper analysis a lot of factors should be taken into consideration, such as: number of tables or seats, number of staff available, location of the restaurant (if it is located in a populated area or in a isolated one), what products it has to offer to its customers (this will contribute to a restaurant seasonality), but unfortunately we only have the daily sales. Once it comes to the daily sales forecasting, the models use information regarding historical values and associated patterns to predict future activity. Most often, this relates to trend analysis, cyclical fluctuation analysis, and issues of seasonality. As with all forecasting methods, success is not guaranteed. The models that could be used in order to lead an accurate analysis are many: the Box-Jenkins Model, for instance, is a technique designed to forecast data ranges based on inputs from a specified time series. It forecasts data using three principles, autoregression, differencing, and moving averages. These three principles are known as p, d, and q respectively. Each principle is used in the Box-Jenkins analysis and together they are collectively shown as an autoregressive integrated moving average, or ARIMA (p, d, q).

II. AIM OF THE ANALYSIS

The aim of the analysis is to investigate daily sales over time, inspecting historical data, examining also their components (-i.e. trend, seasonality and random) concerning six restaurants, located in North of Italy in order to understand in which holidays these restaurants have to restock more their warehouses, when is the best time to open a new location, to get a better idea of the restaurant potential profitability, etc. This will help managers or owners to make decisions with the purpose of increasing revenue and avoiding economic loss. Moreover, diving deeper this research improves owners' decision making in two ways: on one hand, they will know with a certain probability how to allocate resources, for example staff and goods; or from which convenient service the holder should benefit. On the other hand, it will also enable the restaurant manager to warn whether opening the restaurant during a certain holiday can lead to a waste of money or not.

III. METHODOLOGICAL ASPECTS

In order to explore our topic, focusing on the daily sales per each restaurant, we have used different methodological approaches and tools. The motivation behind the usage of certain models over others will be clarified and explained later in the Analysing Data section.

First of all, our deepening into the analysis was merely data-driven, in fact we have let data speak for themselves. Using *matplotilib* library for creating static, animated, and interactive visualizations in Python, exploiting the enormous quantity of methods that this library offer, we took a first look to the main statistics of the data, such as the yearly mean sales per restaurant or the yearly value of the sales during the holidays per each restaurants etc., conducting a first descriptive analytics of data that led us to the predictive analytics. The method used to proceed in this part differs from restaurant in restaurant, because we have considered each of them as single and unique time series. In describing these time series, we have used words such as "trend" and "seasonal" which need to be defined more carefully [2].

- A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear, in fact it could have a "changing direction" when it goes from an increasing trend to a decreasing trend.
- A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency.
- A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. These fluctuations are usually due to economic conditions, and are often related to the "business cycle".

To deepen in the forecasting and to find the best model to fit data, given the nature of the latter that we're going to clarify in the Analysing Data section, we considered these models:

1) SARIMAX;
2) TBATS;
3) PROPHET.

Let's formalize each of it separately.

A. SARIMAX

The abbreviation stands for Seasonal Autoregressive Integrated Moving Average with external variables (SARIMAX) and this model tries to account all the effects due to the sales influencing factors, to forecast the daily sales of each restaurant. In order to understand how this works, we need to introduce the ARIMA model (AutoRegressive Integrated Moving Average), a class of statistical models used to analyse stationary time series (stationarity could be strong or weak, the first requires the shift-invariance (in time) of the finite-dimensional distributions of a stochastic process, while the second only requires the shift-invariance (in time) of the first moment and the cross moment (the auto-covariance)).

This acronym encloses all the main parts of the model:

- AR (AutoRegressive): it uses the dependency relationship among an observation and an n-lagged observations;
- I (Integrated): it makes a series stationary through the differencing of it.
- MA (Moving Average): it uses the dependency relationship among an observation and a residual error through a moving average model applied to the lagged observations.

An Integrated ARMA model of order d is a stochastic process that becomes stationary after differentiating it d times. Since all stationary process can be cast into ARMA(p,q) representations [1], using the AR and MA polynomials in the backward operator B, we have that any integrated process obeys an equation of the kind:

$$\Phi(B)(\mathbf{1}-B)^d Y_t = \Psi(B)\epsilon_t = ARIMA(p,d,q) \tag{1}$$

In real time, many time series have seasonal patterns, in fact it can be introduced seasonal (and non-seasonal) trends, defining a new model called SARIMA(p,d,q)×(P,D,Q) where p,d,q are the autoregressive, difference and moving average orders of the non seasonal polynomials and P,D,Q are the analogous orders for the seasonal part:

$$\Phi(B)\Phi_s(B^s)(\mathbf{1}-B)^d(\mathbf{1}-B^s)^D Y_t = \Psi(B)\Psi_s(B^s)\epsilon_t \tag{2}$$

Going deeper it could be added a X to SARIMA, leading to SARIMAX (The Seasonal Autoregressive Integrated Moving Average eXogenous) models (p,d,q)x(P,D,Q,s) where s is the length of each season -i.e the number of periods necessary to pass before the tendency reappears. For example a SARIMAX model [3] of order (1,0,1) and a seasonal order (2,0,1,5) looks as follow:

$$\begin{aligned} y_t = c + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \\ + \Phi_1(y_{t-5} + \phi_1 y_{t-6}) + \Phi_2(y_{t-10} + \phi_1 y_{t-11}) + \\ + \Theta_1(\epsilon_{t-5} + \theta_1 \epsilon_{t-6} + \theta_2 \epsilon_{t-7}) + \epsilon_t \end{aligned} \tag{3}$$

The total number of coefficients we are estimating equals the sum of seasonal and non-seasonal AR and MA orders. In other words, we're looking at a total of "P plus Q, plus, p plus q" – many coefficients. Furthermore the non-seasonal ones are expressed with lower-case ϕ and θ; while their seasonal counterparts are expressed with upper-case Φ and Θ respectively. Just like with the orders, the capital letters denote the seasonal components and the lower-case ones - the non-seasonal.

B. TBATS

This acronym has in it all the key features of the models:

- **T**rigonometric seasonality;
- **B**ox-Cox transformation;
- **A**RMA errors;
- **T**rend;
- **S**easonal components;

furthermore, it takes its roots in exponential smoothing methods and can be described by the following equations:

$$\begin{aligned} y_t^{(\lambda)} &= l_{t-1} + \psi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t \\ l_t &= l_{t-1} + \psi b_{t-1} + \alpha d_t \\ b_t &= \psi b_{t-1} + \beta d_t \\ d_t &= \sum_{i=1}^{p} \phi_i d_{t-i} \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \end{aligned} \tag{4}$$

where:

- $y_t^{(\lambda)}$ is the time series at moment t (Box-Cox tranfsormed)
- $s_t^{(i)}$ is the i-th seasonal component
- l_t local level
- b_t trend with damping
- d_t is the ARMA(p,q) process for residuals
- ϵ_t is the Gaussian white noise

Each seasonality is modeled by a trigonometric representation based on Fourier series [4]. It requires only 2 seed states regardless of the length of period and it has the ability to model seasonal effects of non-integer lengths. For example, given a series of daily observations, one can model leap years with a season of length 365.25. TBATS will consider various alternatives and fit quite a few models [5]:

1) with Box-Cox transformation and without it;
2) with and without Trend;
3) with and without Trend Damping;
4) with and without ARMA(p,q) process used to model residuals;
5) non-seasonal model;
6) various amounts of harmonics used to model seasonal effects

C. PROPHET

This tool is very powerful and really intuitive, in fact it doesn't require many technical skills. It is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects, that is the main objective of our analysis. It works best with time series that have strong seasonal effects and several seasons of historical data and it is robust to missing data and shifts in the trend, and typically handles outliers well [6]. Prophet uses a decomposable time series model with three main model components: trend, seasonality, and holidays, combined in:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{5}$$

Here g(t) is the trend function which models non-periodic changes in the value of the time series, s(t) represents periodic changes (e.g., weekly and yearly seasonality), and h(t) represents the effects of holidays which occur on potentially irregular schedules over one or more days. The error term represents any idiosyncratic changes which are not accommodated by the model.

This specification is similar to a generalized additive model (GAM) [7], a class of regression models with potentially non-linear smoothers applied to the regressors. Prophet instead uses only time as a regressor but possibly several linear and non-linear functions of time as components. The trend term $g(t)$ could be non-linear and constant or linear and non-constant. In order to define this functions, specific points, where the rate of growth is modified, must be identified. Considering C the carrying capacity,k the growth rate and m an offset parameter, we have:

1) non-linear constant trend:

$$g(t) = \frac{C(t)}{1 + exp((k + a(t)^T\delta)(t + (m + a(t)^T\gamma)))} \tag{6}$$

2) linear non-constant trend:

$$g(t) = (k + a(t)^T\delta)\,t + (m + a(t)^T\gamma) \tag{7}$$

The uncertainty in the prediction of future trend is measured by assuming that the rate of change in an interval T of time of future will have the same frequency and mean numerosity of a past interval T. For the purpose of defining the seasonal components, periodic functions have been used. Considering a standard series of Fourier, the $2N$ parameters are estimated through the construction of a matrix $X(t)$ of seasonal vectors for each instant t for past data and future:

$$\beta = [a_1, b_1, .., a_N, b_N]^T \tag{8}$$

Assuming β normally distributed, we have:

$$s(t) = X(t)\beta \tag{9}$$

Recurring holidays are treated as mutually independent. We ae asked to provide a list of dates considered as holidays and using an indicator function $Z(t)$, Prophet associates each instant t with the change in the forecast relating to the event in question. Assuming k normally distributed, it results:

$$h(t) = Z(t)k \tag{10}$$

This model provides a number of practical advantages:

- Flexibility: It can be easily accommodate seasonality with multiple periods and let us make different assumptions about trends.
- Unlike with ARIMA models, the measurements do not need to be regularly spaced, and we do not need to interpolate missing values e.g. from removing outliers.
- Fitting is very fast.
- The forecasting model has easily interpretable parameters that can be changed by us to impose assumptions on the forecast.

IV. DESCRIPTION OF DATA

The dataset has been provided by university of Milan-Bicocca and it shows the sales figures of six different exercises in the North of Italy's catering sector. Each exercise provides daily sales in euro and the number of receipts:

- **Date**: represents the daily date;
- **Vendite_1**: amount of sales of restaurant 1;
- **Scontrini_1**: number of receipts of restaurant 1;
- **Vendite_2**: amount of sales of restaurant 2;
- **Scontrini_2**: number of receipts of restaurant 2;
- **Vendite_3**: amount of sales of restaurant 3;
- **Scontrini_3**: number of receipts of restaurant 3;
- **Vendite_4**: amount of sales of restaurant 4;
- **Scontrini_4**: number of receipts of restaurant 4;
- **Vendite_5**: amount of sales of restaurant 5;
- **Scontrini_5**: number of receipts of restaurant 5;
- **Vendite_6**: amount of sales of restaurant 6;
- **Scontrini_6**: number of receipts of restaurant 6;

In a preliminary analysis it has been detected the correlation between sales and number of receipts per restaurant.

Fig. 1. Correlation among Restaurants Sales and Receipts

Figure 1 shows that the correlation between sales and receipts is high. In order to avoid the phenomenon of multicollinearity, the focus of the analysis has been maintained only on the daily sales. The data starts from January 2017, but some exercises have been opened after the beginning of the research. That's the reason why, for the latter, the chronicle history has been started later. The lack of availability in terms of data from some restaurants has compromised the precision in the models' estimation.

A. DATA PRE-PROCESSING

Since the goal of our analysis is to investigate during which festivity a restaurant has high sales and therefore whether a restaurant should refuel with goods or not, we have decided to make some changes in the dataset provided initially. These modifications consist in adding new columns:

- *week_day*: it indicates the day of the week;
- *Holidays*: it indicates the holiday associated to that specific day.

in order to analyze the weekly seasonality and the different holidays. The procedure of combining the festivities to the original dataset was possible thanks to a Python library called *holidays*: a fast, efficient Python library for generating country, province and state specific sets of holidays on the fly. It aims to make determining whether a specific date is a holiday as fast and flexible as possible. Concerning the missing values, they were replaced by 0. The projectual choice that stands behind is that if the exercise for that particular day present a missing values, we assume that the sales is equal to 0, in other words the revenue is zero. The updated dataset therefore contains these other two variables:

- **week_day**: represents the day of the week;
- **Holidays**: represents the Holidays days.

V. ANALYSING DATA

In order to lead a smoother study and to avoid redundancy in data and plots, we have decided to search for similar patterns among restaurants carrying out a clustering analysis. Using the free software machine learning library for the Python programming language *scikit-learn 0.24.2* on Jupyter Notebook, we have applied a K-Means Algorithm to the transpose database (with dates as columns and restaurants as index), trying to understand deeper connections among them. Kmeans algorithm [8] is an iterative algorithm that tries to partition the dataset into K-pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within

the same cluster. The objective function, considering 2 clusters, is:

$$J = \sum_{i=1}^{2} \sum_{k=1}^{2} w_{ik} \left\| x^i - \mu_k \right\|^2 \tag{11}$$

where $w_{ik} = 1$ for data point xi if it belongs to cluster k; otherwise, $w_{ik} = 0$. Also, μ_k is the centroid of x^i's cluster. The restaurants are classified in a way that the restaurants 1,2, and 5 are in the Cluster 1, while the restaurants 3, 4 and 6 are in the Cluster 0 (see below for the table).

Restaurants	Cluster
1	1
2	1
3	0
4	0
5	1
6	0

Under these circumstances, we came to a conclusion that we would have analysed just one restaurant per each cluster. Given that, we chose restaurant 1 for Cluster 1 and restaurant 6 for Cluster 0.

A. Descriptive Analysis

Considering all the Sales variables in the dataset per restaurant, we plotted the correlation among them in order to have a deeper understanding of the pattern and clusters. As figure 2 shows, restaurant 1 looks very correlated with restaurant 2 and 5, in fact the trend in the scatterplot follows a straight line. While the restaurant 6 shows an high correlation with restaurant 4 and 3. All the variables seems to follow a normal distribution (despite the 0-values added in the preprocessing part).

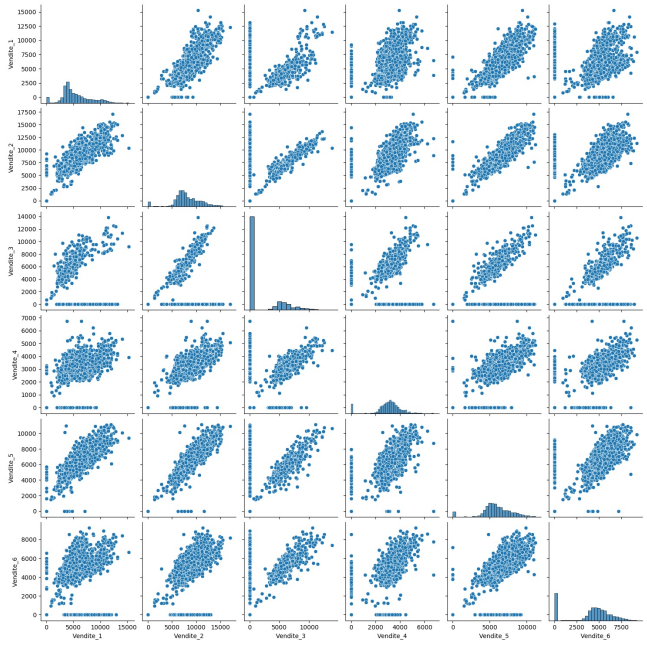

Fig. 2. Correlations among Restaurants' Sales

Furthermore, given that our goal was to understand the behaviours of sales during holidays, we decided to plot the yearly value of sales per each restaurant in specified holidays. In order to avoid redundancy, just holidays such Immaculate Conception and New Years' Eve are plotted. We have grouped the restaurants according to the division given by the clustering analysis, in other words: restaurants 1, 2 and 5 on one side, and restaurants 3, 4 and 6 on the other. The line charts present on x-axis the date of the holiday taken into account (in this case 1st January and 8th December) for each year from 2017 to 2020 or 2021, depending on the festivity. We can notice some salient points: concerning New Year Eve's, the trend experienced a significant decline in 2021, due to Covid-19 restrictions, while restaurant 3 sales in 2017, 2018 and 2019 are null. Overall we can state that with regard to New Year Eve's, all the restaurants trend has decreased.

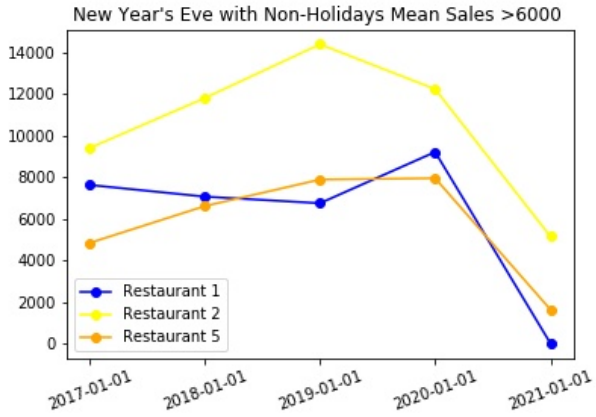

Fig. 3. Sales during New Years' Eve

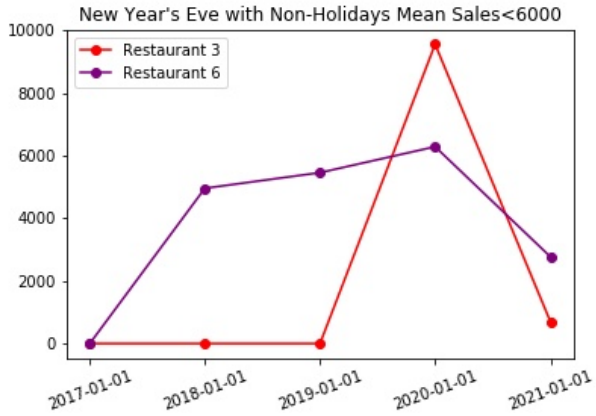

Fig. 4. Sales during New Years' Eve

About Immaculate Conception, we can observe the same trend that fall considerably for all the exercises, also in this case, restaurant 3 was closed on 8th December both years 2017 an 2018 since the sales computed equal 0 euros.

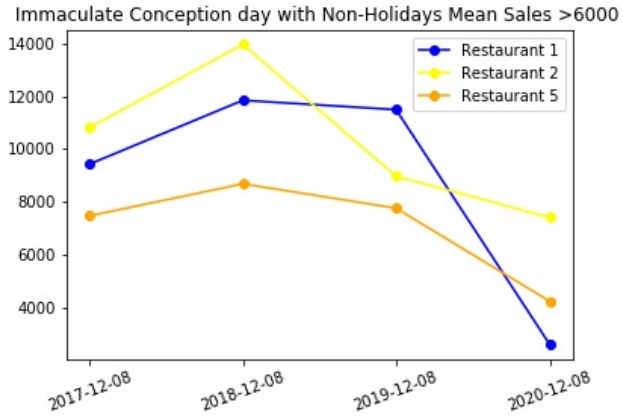

Fig. 5. Sales during Immaculate Conception

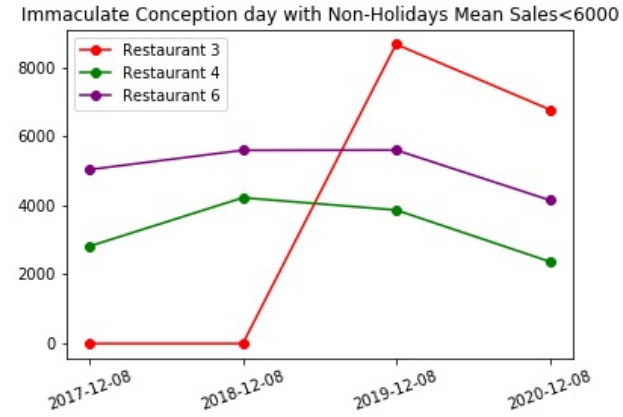

Fig. 6. Sales during Immaculate Conception - part 2

The boxplot in figure 7 gives information about distribution of sales during the week. It is worth noticing that the sales vary from 4000 to 10000 euros. On Monday, Tuesday, Wednesday and Thursday the median sales for the first restaurant is 4000 euros circa but during some week days (for example Monday and Tuesday) the third quartile has reached 5000 euros. Moreover, during the weekend the median sales for restaurant 1 accounted for about 10000 and 8000 euros for respectively Saturday and Sunday. Furthermore, we can also notice the presence of outliers. These are observations that falls below the two whiskers - i.e $Q_1 - 1.5IQR$ and above $Q_3 + 1.5IQR$ and they're represented in the graph by white dots. With IQR, we mean *InterQuartile Range*:

$$IQR = Q_3 - Q_1 \tag{12}$$

in other words the difference between the third (Q_3) and the first (Q_1) quartile. Investigating the distributions, the distribution that described Monday, Tuesday, Wednesday and Thursday sales is positive asymmetric because the difference between the third quartile and the median is greater than the difference between the median and the first quartile; instead for Friday, Saturday and Sunday the distribution is negative asymmetric since the difference between the first quartile and the median is smaller than the difference between the median and the third quartile.

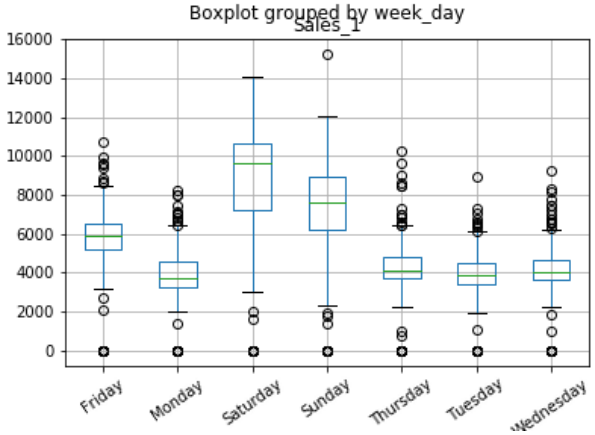

Fig. 7. Distribution of Sales during week days

The boxplots in the following figure allow us to obtain various information regarding the distribution of sales made during the months. It can be observed how sales vary from 2000 to 8000 euro. In general, for the first restaurant, the median of sales over the months varies between about 4000 and 6000 euros, the months with the highest average sales are December, August and September. The presence of upper outliers can also be observed in the months such as February, March, May and July. It is also specified how almost all the outliers are above $Q_3 + 1.5IQR$. Analyzing the boxplots it can also be observed that all of them represent a postive asymmetry, because the difference between the third quartile and the median is greater than the difference between the median and the first quartile.

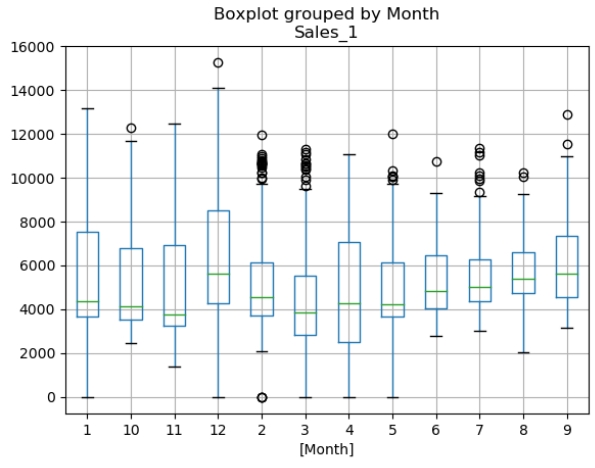

Fig. 8. Distribution of Sales monthly

The boxplot in figure 9 provides information about distribution of sales during 5 years, from 2017 to 2021. In

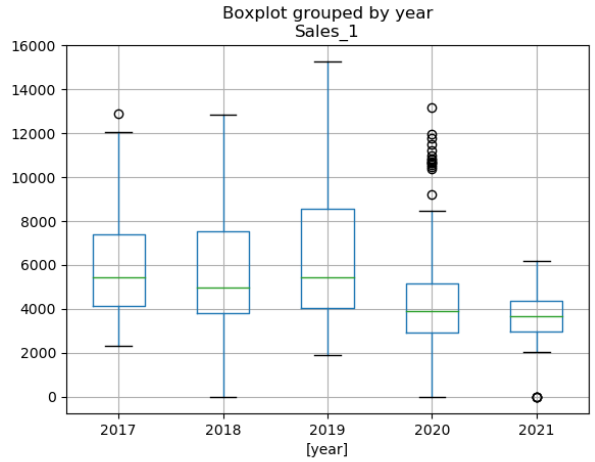

Fig. 9. Distribution of Sales per year

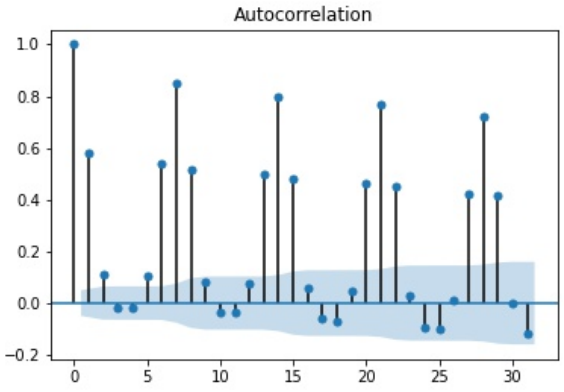

Fig. 10. ACF Restaurant 1

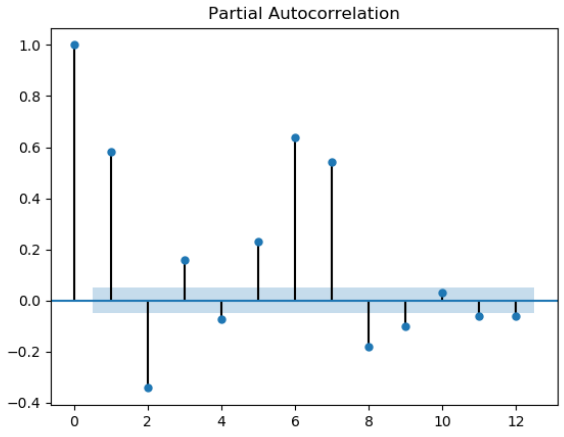

Fig. 11. PACF Restaurant 1

2017, 2018, 2019 the median sales for the first restaurant is between 5000 and 6000 euros, while in 2020 and 2021 it's lower and about 4000 euros. We can notice that in 2020 sales vary from about 3000 to 5000 euros and in 2021 from 3000 to a little bit more than 4000 so the third quartile for both these years is lower than the median sales of the others. In 2020 we can also observed the presence of upper outliers. The distribution that described 2017, 2018, 2019, 2020 sales is positive asymmetric because the difference between the third quartile and the median is greater than the difference between the median and the first quartile; we can't notice any particular asymmetry for the sales in 2021.

Successively, auto-correlation and partial auto-correlation plots were plotted in figure 10 and 11. An auto-correlation plot is designed to show whether the elements of a time series are positively correlated, negatively correlated, or independent of each other. (The prefix auto means "self"— auto-correlation specifically refers to correlation among the elements of a time series). An auto-correlation plot shows the value of the auto-correlation function (acf) on the vertical axis. The horizontal axis of an auto-correlation plot shows the size of the lag between the elements of the time series. For example, the auto-correlation with lag 2 is the correlation between the time series elements and the corresponding elements that were observed two time periods earlier. Auto-correlation with lag zero always equals 1, because this represents the auto-correlation between each term and itself. We can see in figure 10 that there is positive correlation from 5th lag to the 8th lag. These correlations repeat over the lag. That is because each spike that rises above or falls below the dashed lines is considered to be statistically significant. If a spike is significantly different from zero, that is evidence of auto-correlation. A spike that's close to zero is evidence against auto-correlation.

B. Cluster 1: Restaurant 1

First of all, let's look a the line plot of the Restaurant 1 series in figure 12 We could guess, just looking at the graph, that we have 0-values in correspondence of lockdown (light red thick line), and the black line are at holidays. In some holidays there are some daily sales peaks, while in some others not. The daily sales after the lockdown drastically dropped, changing the previous path followed by the series. A moving average is a technique to get an overall idea of the trends in a data set; it is an average of any subset of numbers and it is extremely useful for forecasting long-term trends. We have calculated for a 7-day period of time. Plotting rolling means and variances is a first good way to visually inspect our series. Considering that the rolling statistics exhibit a clear trend (upwards or downwards) and show varying variance (increasing or decreasing amplitude), we can conclude that the series is very likely not to be stationary.

Considering the Restaurant 1 we used the Augmented Dickey-Fuller test, in order to test for a unit root in a univariate process in the presence of serial correlation. As

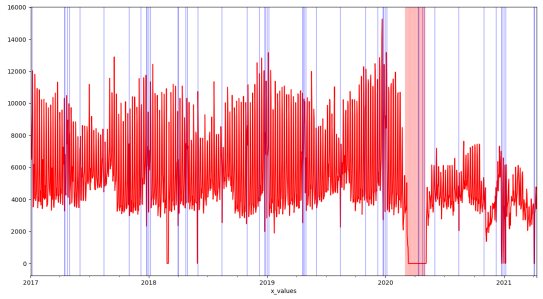

Fig. 12. Raw Data with holidays in black e Covid-19 lockdown in light red of Restaurant 1

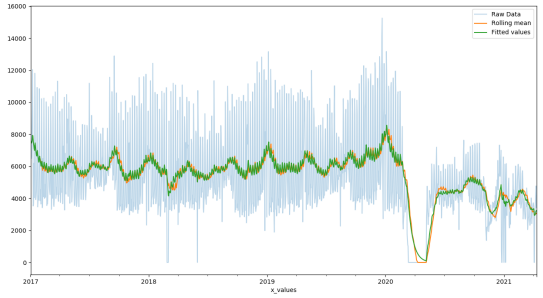

Fig. 13. Rolling mean of Restaurant 1

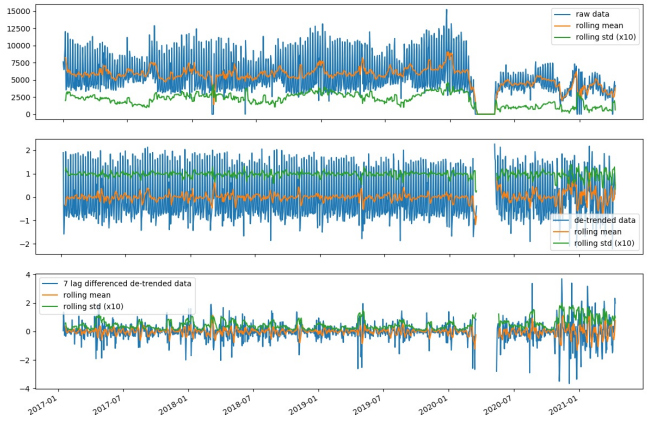

Fig. 14. Stationarity Restaurant 1

Test statistic	P-value	Critical values
-3.324	0.014	1%: -3.43; 5%: -2.86; 10%: -2.57

TABLE I
RAW DATA

Test statistic	P-value	Critical values
-8.668	0.000	1%: -3.43; 5%: -2.86; 10%: -2.57

TABLE II
DE-TRENDED DATA

Test statistic	P-value	Critical values
-8.677	0.000	1%: -3.43; 5%: -2.86; 10%: -2.57

TABLE III
7-LAG DIFFERENCED DE-TRENDED DATA

shown in the table I, the raw data are not stationary with 99% confidence. Furthermore we have also considered de-trended data, computed using the rolling mean and rolling standard deviation:

$$Z_{data} = \frac{x - \mu_{rolling}}{\sigma_{rolling}} \tag{13}$$

Using the Augmented Dickey-Fuller test in table II and looking at the test statistic and p-value, we concluded that de-trended series is likely to be stationary at 99% confidence. Finally, we computed the 7-lag differenced de-trended data:

$$ZP_{7lag} = Z_{data} - Z_{data,7} \tag{14}$$

that is the difference between the de-trended data and the 7-lag de-trended data. Using the Augmented Dickey-Fuller test in table III and looking at the test statistic and p-value, we concluded that the 7-lag differenced de-trended series is likely to be stationary at 99% confidence. As mentioned above, this could also be noted in the figure 14, but the behaviour showed it doesn't seem to stationary, especially for the 7-lag differenced de-trended data.

A lag plot is a special type of scatter plot with the two variables "lagged"; a "lag" is a fixed amount of passing time. One set of observations in a time series is plotted (lagged) against a second, later set of data. The lag plot in figure 15 checks whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot; the points in the lag plot appear scattered from left to right and top to bottom. Non-random structure in the lag plot indicates that the underlying data are not random; data with auto-correlation gives rise to lag plots with linear patterns that follow the diagonal. As the level of auto-correlation increases, the points cluster more tightly along the diagonal.

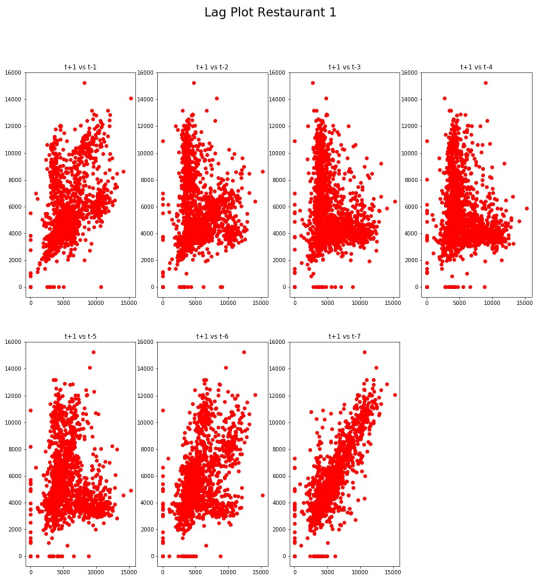

Fig. 15. Lag Plot Restaurant 1

Lag plots can also help to identify outliers. We can notice a linear pattern only in the 7th order lag plot indicates probable auto-correlation.

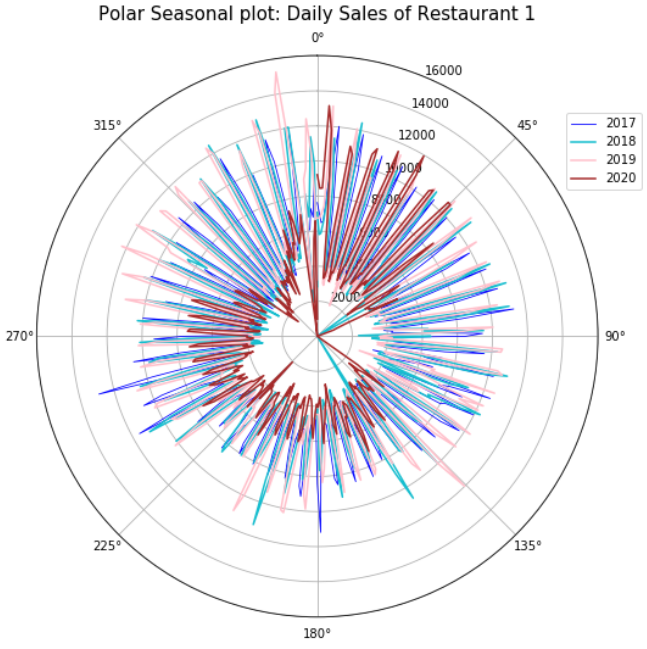

Fig. 16. Polar Seasonal plot for Restaurant 1

A seasonal plot is similar to a time plot except that the data are plotted against the individual "seasons" in which the data were observed. The polar seasonal plot is a variation of seasonal plot, because it uses the polar coordinates represented by days of the year. Figure 16 shows us the strong presence of weekly seasonality in first restaurant's sales, highlighting the peaks during the weekends.The daily sales after the lockdown started again following the weekly seasonality, however reducing the mean value.

1) Trying models: In order to find a model that can describe or,in other words, fit our data, we've modelled data according to several models. Firstly, the dataset was split into 2 subsets: the test and the train set consisting in respectively in 30 days and $1563-30=1533$ days in order to build the model on the training dataset and forecast using the test dataset, (1563 is the total number of observations for restaurant 1). The projectual choice behind the fact of having such a huge train set stands for the issue of including in the training set also the period of time characterized by Covid-19 spread. The method *auto_arima* is contained in Pmdarima: a statistical library designed to fill the void in Python's time series analysis capabilities. This includes:

- the equivalent of R's auto.arima functionality
- time series utilities, such as differencing and inverse differencing
- a collection of statistical tests of stationarity and seasonality
- numerous endogenous and exogenous transformers and featurizers, including Box-Cox and Fourier transformations
- seasonal time series decompositions
- cross-validation utilities

The auto-ARIMA process seeks to identify the most optimal parameters for an ARIMA model, settling on a single fitted ARIMA model. Auto-ARIMA works by conducting differencing tests to determine the order of differencing, d, and then fitting models within ranges of defined start_p, max_p, start_q, max_q ranges. If the seasonal optional is enabled, auto-ARIMA also seeks to identify the optimal hyper-parameters P, Q and D.

In order to find the best model, auto-ARIMA optimizes for a given information_criterion that can be one of 'aic', 'aicc', 'bic', 'hqic', 'oob' which correspond respectively to Akaike Information Criterion, Corrected Akaike Information Criterion, Bayesian Information Criterion, Hannan-Quinn Information Criterion. The best ARIMA model is the one which minimizes one of these values. It's worth pointing out that due to stationarity issues, auto-ARIMA might not find a suitable model that will converge.

Initially, the model studied is SARIMAX. Firstly, the needed libraries were imported, subsequently the model was trained in order to forecast future data. To construct the Sarimax model, we have taken into account the parameters provided by *auto_arima* method; therefore, these parameters were passed in order to train the model used to forecast the data, more precisely to forecast 30 days. In fact in figure 17, it is worth noticing that the Sarimax model fit very well comparing to the other models studied -i.e. tbats and prophet; in fact taking a look on RMSE (Root Mean Square Error) in table IV the value associated to the Sarimax model is the smallest one; in fact the RMSE

value for restaurant 1 is equal to 950.05. It is important to point out that we've forecasted the past using Sarimax model, this technique was used in order to evaluate the model's accuracy. Indeed, the model obtained is able to recognize the trend that characterised the time series for restaurant 1. When the test set present an increase, the Sarimax follow the increasing trend, on the other hand we can notice the same exact behavior when there is a decrease. For example from 15th to 22th of March, the model detect in a reliable way the data. Over time, the model's accuracy decrease due to the fact that as time passes, the estimate of the forecast become less precise.

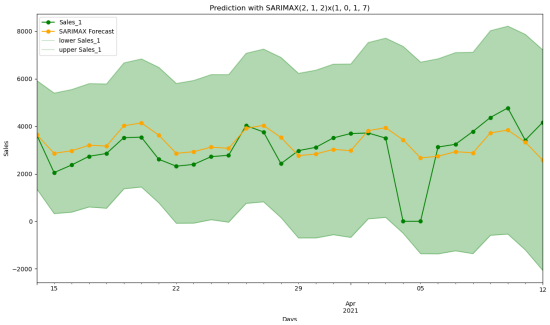

Fig. 17. Sarimax(2,1,2)x(1,0,[1],7) Forecasting vs Test data of Restaurant 1

Specifically for our model, Sarimax(2,1,2)x(1,0,[1],7) means that it is describing some response variable (Y) by combining a 2nd order Auto-Regressive model and a 2nd order Moving Average model, and it means thata was differenciated only once; regarding seasonal elements, the fitted model has combined a 1st Seasonal auto-regressive order model and a 1st Seasonal moving average order. Moreover the last value presented indicates the number of time steps for a single seasonal period. Looking at graph 17 we can notice that the 95% confidence interval grows over time and this suggests us that this model is useful in the short forecasting period in order to have more precision. Furthermore, not considering the holidays it couldn't predict the decrease on 4th of April, Easter Day in Italy.

Finally we've used the library PROPHET to forecast the future sales. As previously seen, the dataset was split into train and test set. As we can notice the blue line reports the restaurant's sales, while the yellow line represents the model's prediction. In addiction the purple and red lines set the confidence interval of the prediction with an alpha equal to 0,5. The purple line is the lower bound, while the red one represents the upper bound.

In figure 19 the effects, such as trend, holidays, weekly seasonality and yearly seasonality, of the time series components are plotted. It's worth to notice that the holidays' effect has a large impact on the daily sale of the restaurant, denoting losses on some holidays and gains on others. For example on New Years' Eve and Easter of every year, the losses reach a value of about −4000

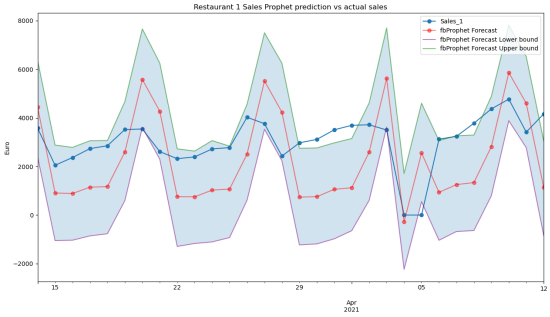

Fig. 18. PROPHET Forecasting vs Test data of Restaurant 1

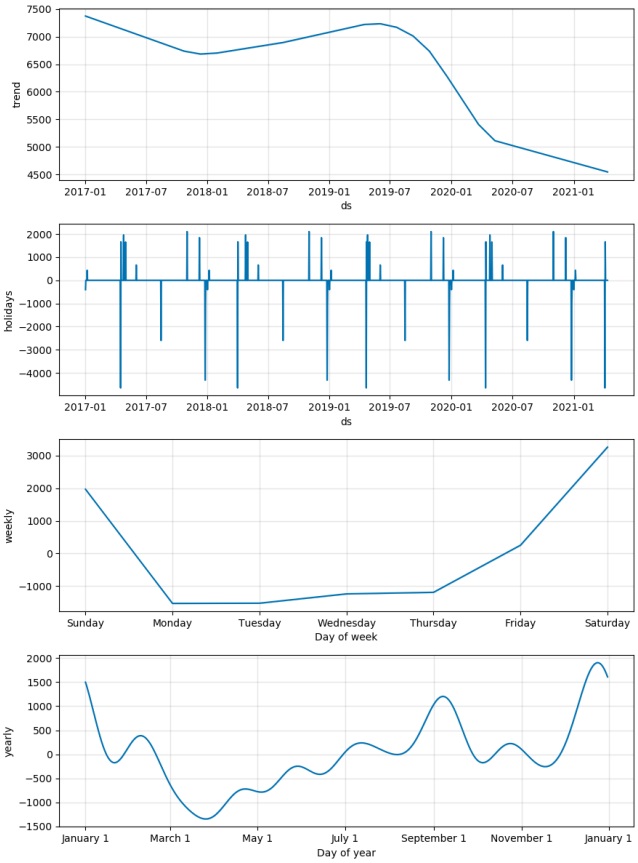

Fig. 19. Effects of Components on the series

euro, suggesting that the restaurant should be closed rather than opened. Considering instead holiday such as "Immaculate Conception", it can be noticed that there is a gain every year, of about +2000 euro, increasing the daily sale. However, the effect that has the greatest impact on sales is the weekly seasonality. The graph shows that when there are weekend days, the sales increase a lot more than during week days, in particular on Friday and Saturday. Yearly seasonality also has an impact, showing that on December and October the sales are greater than in other months. As it can be noticed in figure 20, the most important effect that has the greater impact of daily

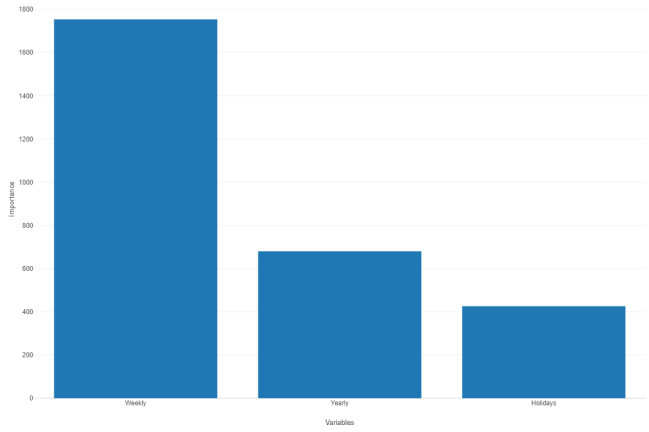

Fig. 20. Importance of components in the sales

sales is the weekly seasonality.

C. Cluster 0: Restaurant 6

In order to make the paper more streamlined, we decided not to report the stationarity analysis for restaurant 6 because it was very similar to that of the first group of restaurants. With this we decided to exclusively report the testing of the various models after eliminating the first 9 months of 2017 as the values were missing. Also to test each model we have divided into train and test set, where the latter has a size of 30 days (as for the first group of restaurants). The dataset of the daily sales of restaurant 6 has the values of the first months of 2017 missing, considering that we've decided to associate 0 to every missing values, looking at graph 21, that shows the line chart of restaurant 6, we can notice what mentioned above. Considering this, in order to not compromise the training of the models, we decided to remove these first months from the train set.

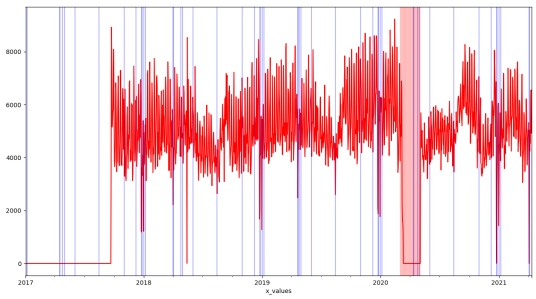

Fig. 21. Raw Data with holidays in black e Covid-19 lockdown in light red of Restaurant 6

Overall we can notice that there are not some visible trends and seasonality, and that after the lockdown, highlighted in red in the graph, caused by Covid 19, the daily sales are slightly decreased.

As we have done for Restaurant 1, we have developed different models in order to find the ones that may describe in a proper way the performance of the sixth restaurant. The first model presented in figure 22 is Sarimax; as previously seen, we have taken into account the parameters provided by *auto_arima* method; after, these parameters were passed in order to train the model. Looking deeper in figure 22, we can notice that the model was able to forecast performances for a certain period of time, for example the last week of March.

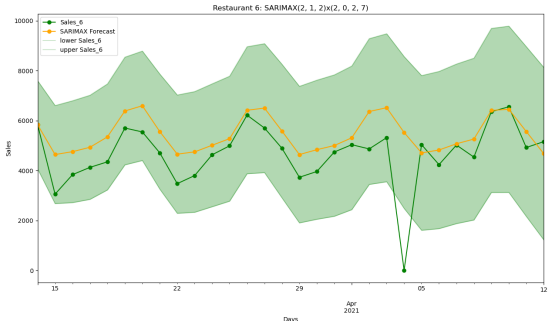

Fig. 22. Sarimax Forecasting vs Test data of Restaurant 6

Considering the plot in figure 23 where are the values of the test set plotted against the values predicted by Prophet, it can be noticed that this model has also predicted the decrease of sales in April. This was accomplished thanks to the Holiday variable, in fact on 4th of April 2021 falls the Easter day.

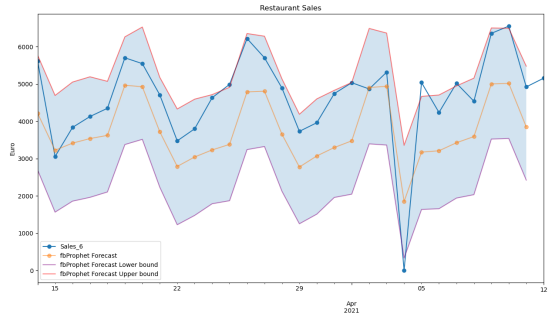

Fig. 23. PROPHET Forecasting vs Test data of Restaurant 6

The graph in figure 24 are the Prophet components and the Holiday component that show some increases and decreases corresponding to the Holidays during the year. As we expected from the graph 23, every year during the Easter day there's a decrease in the restaurant 6 sales. Furthermore, from Monday to Sunday there is constant growing of the sales, reaching its peak on Friday, Saturday. The months associated to the higher sales are about October, December and February.

VI. RESULTS

In order to choose the best model we have used the Root Mean Squared Error metrics. Considering that for restaurant 6 the model with the smallest RMSE is Prophet,

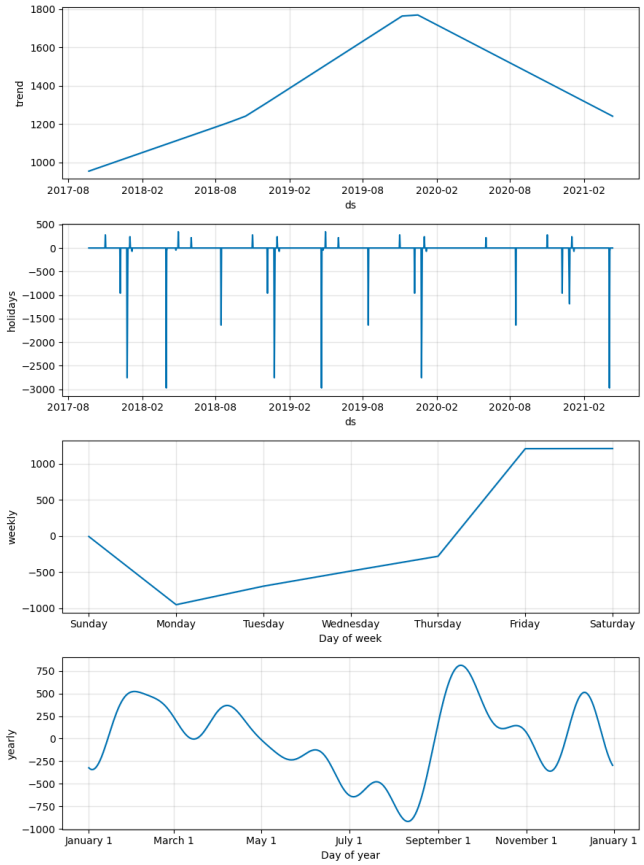

Fig. 24. Effects of Components on the series

we used it to forecast one year of daily sales; for restaurant 1 the best model, with the lowest RMSE was the Sarimax so we used it to make 90 days sales predictions.

Models	RMSE R1	RMSE R6
SARIMAX	950.05	1208.27
TBATS	1064.15	1532.91
PROPHET	1838.44	776.34

TABLE IV
RESULTS OF RMSE'S MODEL

TBATS predictions were not displayed above because neither of restaurants had a good root mean squared error. Considering the first restaurant the significant parameters are summarized in figure 25. The number to look at first in the training summary is the Ljung-Box test's statistic and its p-value. The Ljung-Box helps us determine if the residual errors of regression are auto-correlated in a statistically significant way. In this case, the p-value is 0.6 which is significantly higher than 0.05 (95% confidence threshold). So we accept the null hypothesis of the Ljung-Box test that the residual errors are not auto-correlated.

Two other things to note in the result: The Jarque-Bera test of normality has yielded a vanishingly small p-value implying a rejection of the Null hypothesis at a > 99.99% confidence level. The Null hypothesis being

SARIMAX Results

Dep. Variable:	y	No. Observations:	286
Model:	SARIMAX(2, 1, 2)x(1, 0, [1], 7)	Log Likelihood	-2384.529
Date:	Tue, 20 Jul 2021	AIC	4783.058
Time:	18:13:19	BIC	4808.625
Sample:	0	HQIC	4793.307
	- 286		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8542	0.193	4.428	0.000	0.476	1.232
ar.L2	-0.4425	0.100	-4.416	0.000	-0.639	-0.246
ma.L1	-1.2035	0.202	-5.970	0.000	-1.599	-0.808
ma.L2	0.3438	0.182	1.892	0.059	-0.012	0.700
ar.S.L7	0.9606	0.027	35.087	0.000	0.907	1.014
ma.S.L7	-0.8606	0.055	-15.764	0.000	-0.968	-0.754
sigma2	1.066e+06	5.52e+04	19.326	0.000	9.58e+05	1.17e+06

Ljung-Box (Q):	37.04	Jarque-Bera (JB):	353.61
Prob(Q):	0.60	Prob(JB):	0.00
Heteroskedasticity (H):	1.03	Skew:	-0.62
Prob(H) (two-sided):	0.89	Kurtosis:	8.31

Fig. 25. Sarimax Summary

that the regression errors are normally distributed. This is probably because the errors are highly kurtotic (note that the Kurtosis=8.31 as against the 3.0 that it should have been for a normal distribution). Note also that the errors are not at all skewed (skewness=-0.62 as against 0.0 for normally distributed errors). The overall analysis of the parameters and the tests done show us a good model to use in order to forecast in the short period of time. In fact we decided to forecast 90 days from last recorded sale (12/04/2021), using SARIMAX(2,1,2)x(1,0,1,7). Looking at graph 26, it can be noticed that the 95% confidence interval grows as fast as days go on, in fact larger is the horizon we want to forecast, the worst is the precision in the forecasting. Given that this model is good to have a look in about two months ahead. As showed by the read predicted daily mean, the sales increase over time.

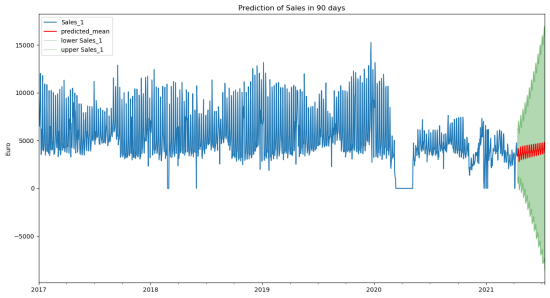

Fig. 26. 90 Days SARIMAX(2,1,2)x(1,0,1,7) Forecasting for Restaurant 1

Considering instead the restaurant 6, the RMSE in table IV revealed that the best model was PROPHET, probably because the 6th restaurant is more influenced by the holidays than the first one. Considering this, we decided to cross validate this model. This has been possible because Prophet [6] includes functionality for time series cross validation to measure forecast error using historical data. This is done by selecting random cutoff points in the history, and for each of them fitting the model using data only up to that cutoff point. We specify the forecast horizon (30 days), and then the size of the initial training period (three times the horizon) and the spacing between cutoff dates (every half a horizon). In order to evaluate the cross validation, we computed the MdAPE - Median Absolute Percentage Error in figure 27. As the graph 27

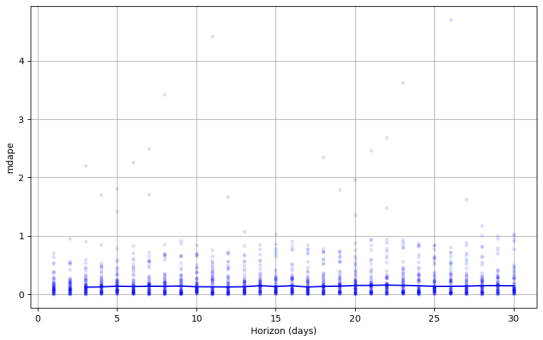

Fig. 27. Cross Validation of Prophet Model - Median Absolute Percentage Error in figure

shows, the MdAPE it's very small during the entire horizon, in fact it stays around 0.1 mostly. This is a good sign, so we decided to use the Prophet model to forecast the 90 days after the last recorded sale on 12/04/2021, in figure 28.

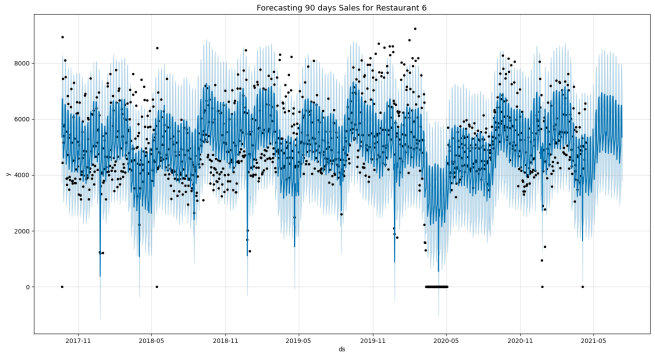

Fig. 28. 90 Days PROPHET Forecasting for Restaurant 6

The light blue part is the 95% confidence interval and it's not increasing overtime as in the forecasting of the first restaurant. It is very useful to notice how the Prophet model is able to forecast the peak and the decreases due to holidays, and also the most earn-able month during the year. There is a noticeable increase when it comes to May 2021, maybe due to 1st of May Holiday and the start of the summer season.

Overall, we can come to the conclusion that the two groups of restaurants are very different between them. First of all, the second one, Cluster 0 that includes the 6th restaurant, is more influenced by the holidays, in fact it has serious loss on some holidays and gains on others. For example, on Easter Days and during August the loss are huge, so we can assume that in the area where it is, people doesn't go out on Easter and it is not a place visited on summer. Considering that in this paper [9] we analysed the usual summer Italian holiday, we can assume that this group of restaurants are neither on the high mountains nor near the sea. On the other hand it has some gains on October and on December (on the 8th of December, Immaculate Conception). Second graph in figure 24 shows that even if in some holidays it gains, the earnings are always under the 500 euro. The first group of restaurants (Cluster 1) instead showed a more seasonal influence of the sales, in fact having a weekly seasonality, it can be noticed that on the weekend the sales increase a lot with respect the weekdays. From Monday to Thursday the sales perceive a loss, so it would be better if it stays close. Considering this we could assume that it is in a small town were people are working hard on weekdays and goes to the restaurants only on weekends.

VII. CONCLUSION AND FURTHER DEVELOPMENTS

First of all, considering that our dataset has both the sales and the receipts variable, we checked the correlation among them to avoid multicollinearity. As a matter of fact we found out that the correlation was steep and we decided to consider just the sales variable. Having divided the restaurants into two clusters, during the analysis we noticed that the first and second group are different among them. The first group of restaurants is more influenced by the weekly seasonality, in fact on the weekend the sales are higher than the weekdays. Because of this, in correspondence of weekends, the owners should call more employees to work and should increase the stocks in the warehouse. The second cluster, instead, faces serious losses or has high revenues according to holidays. For example, on Christmas Days and during April the losses are undue. Considering that during these years there was a black swan event such as Covid-19, we decided to forecast just a short period of time (90 days). A future development of our project could be focused on predicting a longer period of time, maybe 6 months, in order to ease the refuelling of restaurant's owners. Adding more variables to our analysis, such as:

- refuelling costs;
- staff costs;
- management costs;
- customers reviews;
- restaurant's capacity (square meters).

It could be developed a software that installed on owner's devices can act as a management system, in order to improve the catering. Furthermore, a sentiment analysis might be carried out to let owners understand restaurants' pros and cons. Considering these black periods of economic crisis, it is crucial to have by your side a good analyst that could lead you to make more efficient decision, to boost responsible behaviours in a society that lives wasting food.

REFERENCES

[1] Fattore M. (2020).Fundamentals of time series analysis, for the working data scientist (DRAFT).

[2] Hyndman, R.J., & Athanasopoulos, G.(2018). Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. [Accessed: 16-Jul-2021].

[3] 365datascience.com (2021). What Is a SARIMAX Model? [Online]. Available: https://365datascience.com/tutorials/python-tutorials/sarimax/. [Accessed: 16-Jul-2021].

[4] medium.com (2019). Forecasting Time Series with Multiple Seasonalities using TBATS in Python [Online].Available: https://bit.ly/3iGI4ZR. [Accessed: 16-Jul-2021].

[5] De Livera, A.M., Hyndman, R.J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing, Journal of the American Statistical Association, 106(496), 1513–1527. Working Paper version is available at https://robjhyndman.com/papers/ComplexSeasonality.pdf. [Accessed: 16-Jul-2021].

[6] DTaylor SJ, Letham B. 2017. Forecasting at scale. PeerJ Preprints 5:e3190v2 Working Paper version is available at https://doi.org/10.7287/peerj.preprints.3190v2. [Accessed: 16-Jul-2021].

[7] Hastie, T. & Tibshirani, R. (1987), Generalized additive models:some applications, Journal of the American Statistical Association82 (398), 371-386.

[8] towardsdatascience.com (2019), K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Available: https://bit.ly/2WwQ2wY. [Accessed: 11-Jun-2021].

[9] Chabib S., Grasso L., Schiavone A., (2020), Statistical Analysis of Italian Trips. Available: https://github.com/liliagrasso/progettiepubblicazioni/blob/main/Lilia_Grasso_813210_Statistical_Analysis_on_Italian_trips.pdf. [Accessed: 19-Aug-2021].