Data Science Lab

Project – Report

15/06/2023

Yuliia Tsymbal, 894213,

Induni Sandampiumi Nawarathna Pitiyage, 906451,

Yasmin Bouhdada,837389,

Sara Campolattano, 906453

INTRODUCTION

In this report we aim to present a time series analysis of the “restaurants” dataset, carried out with the intention of developing a forecasting model.

After briefly introducing the above-mentioned dataset, we will propose our main objectives and the motivations that led us to formulate such ideas.

In particular, through means of exploratory data analysis, we will show in detail how the target variable, i.e., the total gross earnings (“lordo totale”) of the restaurants changes over time, covering a time span that goes from September 1st, 2018, to May 3rd, 2023, and we will ascertain that the major requirements for dealing with time series data are met.

Moreover, we will show how the information gathered from EDA led us to the decision of enriching the data originally available with additional factors coming from external sources, aiming to have a better understanding of the context, which will allow us to develop an accurate forecasting model.

Last but not least, we will introduce the model developed to forecast the total gross earnings of the restaurants, by means of Prophet forecasting tool, for a future time period of seven days, and we will see how said earnings vary over the given time span.

DATASET

The original dataset around which this project has been developed contains information about six restaurants, namely “R000”, “R001”, “R002”, “R003”, “R004” and “R005”, located in Montebello (PV), Piacenza (PC), Stradella (PV), and Voghera (PV) respectively, and their daily total gross earnings over a time period that goes from January 1st, 2018, to May 31st, 2023, with the exception of restaurant “R002”, as its data was acquired starting from November 1st, 2019.

Specifically, the dataset consists of a total of 11.188 observations and four variables, namely “date”, corresponding to every day in the span of the above-mentioned time period, “scontrini”, referring to the number of receipts per day, “lordo totale”, i.e., the total, gross earnings, and “restaurant”, corresponding to the code of each restaurant.

We care to point out that, for the first eight months, i.e., from the time period that goes from January 1st, 2018, to August 31st, 2018, information regarding “lordo totale” of the restaurants is available in the form of monthly data instead of daily data.

In order to preserve the original structure of the time series data, the analysis proposed in the following sections has been carried out on a subset of the original dataset, consisting of 9.809 observations, covering a time span that goes from September 1st, 2018, to May 3rd, 2023, as it corresponds to a period for which the daily data regarding “lordo totale”, i.e., the total, gross earnings of the restaurants, is available.

MOTIVATIONS

As modern automation systems for data collection have been created, data scientists and researchers are faced with an ever-growing amount of data, often coming in the form of time series.

Time series data represent sequences of observations recorded over time; they enable us to uncover hidden patterns, trends, and dependencies within the time dimension, becoming a crucial tool for decision-making processes and predicting future outcomes.

As previously mentioned, the “restaurant” data covers a time span that goes from September 1st, 2018, to May 3rd, 2023, namely almost five years. The amount of information concerning the total, gross earnings of the six restaurants, residing in this time interval, could be potentially useful to uncover valuable insights and develop a robust forecasting model capable of capturing the inherent complexity of time-varying data.

The aim of this project, therefore, is to provide, firstly, a context overview, in order to understand whether some useful information can be extracted by the data itself; secondly, we aim to build a forecasting model that is able to accurately predict the total, gross earnings, i.e., “lordo totale”, for a future time period of seven days, for each one of the restaurants.

EXPLORATORY DATA ANALYSIS & ENRICHMENT

As previously mentioned, in this section we present an overview of the data, aiming to provide some valuable information for future reference.

First and foremost, as we are dealing with time series data, it is important to ascertain some main checkpoints before actually begin to develop a forecasting model. In order to achieve this task, it is fundamental to understand (1) whether the data is stationary, (2) if seasonality can be detected and (3) whether the target variable, which in our case is "lordo totale", is autocorrelated.

To do so, we start by taking a look at Figure 1, which shows how the daily total gross earnings change over a time span that, as mentioned above, goes from September 1st, 2018, to May 3rd, 2023. We care to note that, for simplicity, we decided to take as reference the restaurant R001, located in Piacenza (PC). The analysis carried out on all the restaurants can be found at the link (1.) in the "References" section.

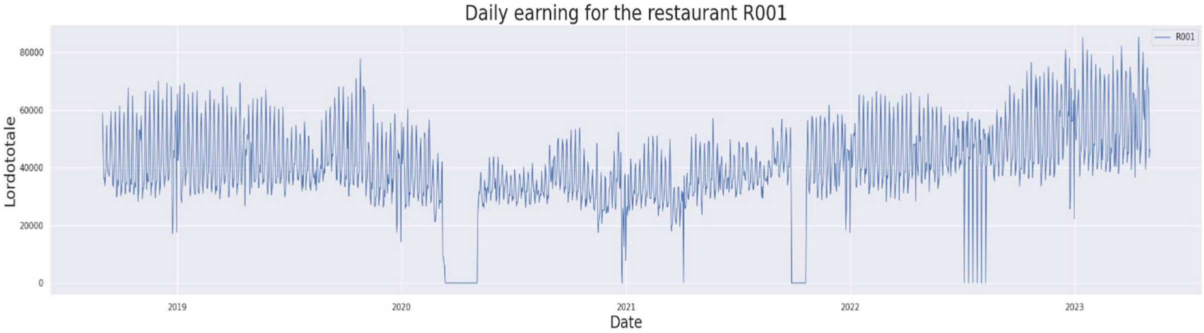

Figure 1

As we can see from Figure 1, it is not possible to detect a specific and clear behavior of "lordo totale", i.e., the total, gross earnings, of restaurant R001 over the given time period. Specifically, we can see that from September 1st, 2018, to March 2020, "lordo totale" seems to oscillate almost homogeneously. However, from March 2020 we can observe an outrageously steep decline in the total, gross earnings, bringing its value to zero. Around June 2020, we can observe a sudden increase in "lordo totale", which behavior appears to remain approximately stable until the end of the year. As a matter of fact, at the end of 2020 we can see another sharp decrease in "lordo totale", which, however, appears to concern only a day. It is from this point in time that we can observe an unstable behavior of the total, gross earnings of restaurant R001, with random steep negative peaks, and progressive increases simultaneously, till May 2023.

The information gathered up to now, led us to believe that (1) there must be some external factors causing the total, gross earnings of the restaurants to drop significantly in 2020 and sporadically from the end of 2020 to 2023, and (2) the daily data regarding the total, gross earnings of the restaurants can be analyzed separately according to three different time intervals (following the above statements concerning Figure 1).

To investigate potential factors causing “lordo totale” to drop in such a manner, we decided to zoom in those precise points in time and search for additional information that might explain said behavior. Our search led us to discover that, according to news articles and ISTAT data, the steep decline in the total, gross earnings from March 9th, 2020, to June 5th, 2020, and from December 25th, 2020, to January 6th, 2021, was due to the Covid-19 Pandemic.
As it played a major role in the behavior of the total, gross earnings of the restaurants, we therefore decided to integrate to our original subset the Covid-19 Pandemic data (2.) pertaining the cities in which the restaurants are located, so as to possibly improve the forecasting modeling. Additionally, to discover what caused “lordo totale” of the restaurants to drop randomly from January 6th, 2021, to May 3rd, 2023, we zoomed in those exact points in time and we discovered that, hypothetically, part of those negative peaks is due to the restaurants’ closing days.

Moreover, to better understand the behavior of the total, gross earnings of the restaurants, and not only its decline, we were able to hypothesize the influence of more than one external factor, that is, the Covid-19 Pandemic.

Considering the context in which we are operating, several are the potential factors that could influence “lordo totale” of the restaurants to both increase and decrease. For instance, it could be correct to affirm that, in general, restaurants tend to earn more during weekends than during weekdays; it could also be correct to hypothesize that they earn more during holidays than workdays, or more in sunny days than rainy days.

For these reasons, we decided to enrich the original data integrating more information: we listed all the weekends, national holidays, “feste patronali”, Ramadan months and the most important football matches (e.g., European Champions Legue, World Cup) for modeling. Moreover, we integrated weather information (3.) for the whole time series span, indexing the natural phenomena as follows: no phenomena (code: 1), grandine-nebbia (code: 2), nebbia (code: 3), neve (code: 4), neve-nebbia (code: 5), pioggia (code: 6), pioggia-nebbia (code: 7), pioggia-neve (code: 8), pioggia-neve-nebbia (code: 9), pioggia-temporale (code: 10), pioggia-temporale-nebbia (code: 11).

As previously stated, following the statements made with regards to Figure 1, we believe that the daily data regarding “lordo totale” of the restaurants can be analyzed separately according to three different time intervals, namely the pre-Covid-19 Pandemic time period, the period during Covid-19 Pandemic, and the period after it. This is due to the fact that the behavior of the total, gross earnings of the restaurants seems to suggest different trends in each one of these periods and, therefore, we considered it would be appropriate to conduct further analyses to detect stationarity, seasonality and autocorrelation separately.

DICKEY-FULLER TEST

Firstly, we would like to start by the concept of stationarity, as it is one of the fundamental characteristics of time series. A time series is said to be stationary if its statistical properties do not change over time; in other words, it has constant mean and variance, and covariance is independent of time. In order to model time series, it is crucial to transform non-stationary data to a stationary process.

To understand whether our time series is stationary, we performed the Dickey-Fuller Test on three different time intervals: the first period going from September 1st, 2018, to March 8th, 2020, the second period going from March 9th, 2020, to December 27th, 2020, the last period going from December 28th, 2020, to May 3rd, 2023.

The Dickey-Fuller test is a statistical significance test, meaning that the test will output results in terms of hypothesis tests, hence, involving null and alternative hypotheses. As a result, said test provides a p-value from which we will need to make inferences about whether the time series is stationary or not. The presence of a unit root in the time series defines the null hypothesis, while the alternative hypothesis defines time series as stationary. Since the null hypothesis assumes the presence of a unit root, the p-value obtained by the test should be less than the significance level (e.g., 0.05) to reject the null hypothesis. Thereby, inferring that the series is stationary.

We care to note that, for simplicity, we decided to take as reference the restaurant R001, located in Piacenza (PC).

After performing the Dickey-Fuller Test on the three different time periods defined above, we obtained the following results:

	First Time Period	***Second Time Period***	***Third Time Period***
	01/09/2018 – 08/03/2020	09/03/2020 – 27/12/2020	28/12/2020 – 03/05/2023
p-value	0.398	0.140	0.015

Table 1

As we can see from the above table, the p-value of the Dickey-Fuller Test is greater than the significance level, set at 0.05, suggesting that the time series is not stationary, as it is not possible to reject the null hypothesis, and that seasonality and autocorrelation can be detected. The p-value of the test, also for the second period is higher than the significance level, meaning that the null hypothesis, that is, there is a unit root in the time series, cannot be rejected. For what concerns the third period, however, we can see that the p-value is slightly lower than the significance level, which leads us to reject the null hypothesis rejected in favor of the alternative hypothesis, i.e., the data is stationary.

As previously mentioned, in order to properly model time series data and to better forecast, stationarity is required. For this reason, to make the time series stationary and to remove underlying autocorrelation issues, we simply subtracted the time series from itself, with a lag of 30 days. This procedure allowed us to obtain fully stationary time series for all the three different periods.

PROPHET FORECASTING TOOL

Prophet is an open-source forecasting library developed by Facebook's Core Data Science team and it was designed to provide accurate and intuitive forecasting of time series data. Prophet uses an additive model, that is, it decomposes time series data into trend, seasonality, and holiday components, as it assumes that future values of a time series can be expressed as the sum of these components. The algorithm takes into account various seasonal patterns present in the data, such as weekly, monthly, or yearly cycles, and it also provides mechanisms to handle missing data and outliers in a principled manner.

To further prove how the behavior of the total, gross earnings of the restaurants changes in the three different time periods, we decided to apply Prophet to understand the differences in terms of trends. We care to highlight that, for simplicity, also in this case we decided to take as reference the restaurant R001, located in Piacenza (PC).

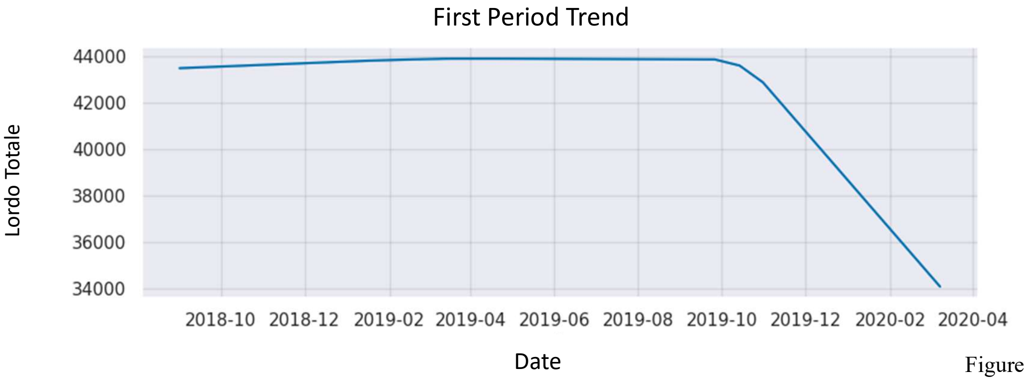

Figure 2

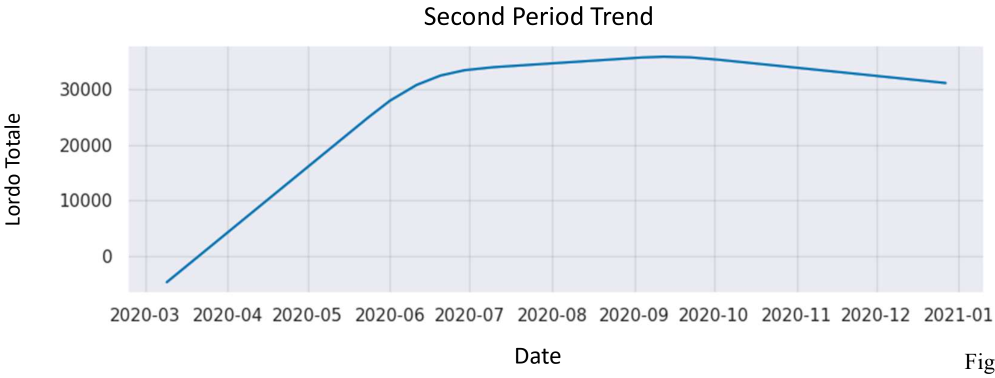

Figure 3

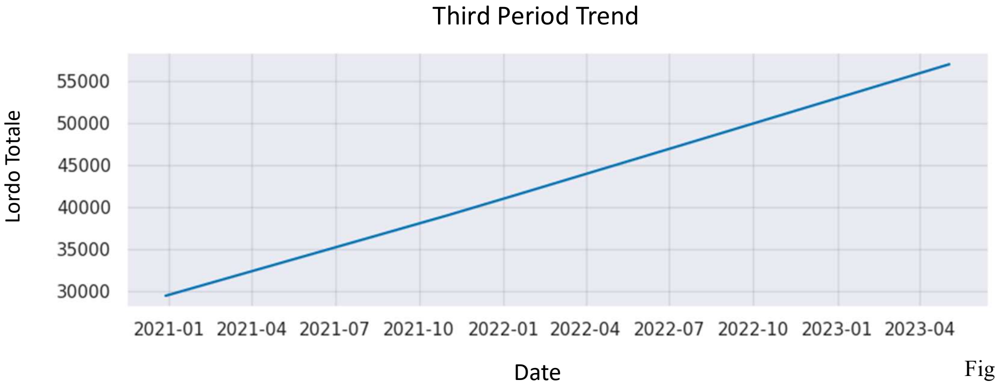

Figure 4

As we can see from Figure 2, “lordo totale” appears to have a steady behavior till October 2019, when it starts decreasing dramatically, reaching its lowest value around March 2020.

This behavior finds its explanations in the starting of the Covid-19 pandemic. In Figure 3, as expected, we can see a rapid increase in the total, gross earnings from March 2020 to June 2020, as June corresponds to the month in which the lockdown interval was dropped, and then an approximately flat behavior for the rest of the year. In Figure 4, in which the third period is represented, we can observe a linear and constant increase in "lordo totale", an increase that covers a time span that goes from the end of 2020 to May 2023.

In general, taking into consideration the whole picture, we can observe how the total, gross earnings of the restaurants drops from 44.000 € to 0 € in the first period, how it rises from 0€ to 30.000 € in the second period, and how it keeps increasing till 55.000 € in the third period.

Therefore, all three periods have a different structure, and the trend is notably different in each of them. Moreover, we discovered that weekly seasonality can be detected in all three periods, but it shows different characteristics. Additionally, the third time period exhibits annual seasonality (although this can be justified given that only this time period has data for two full years).

MODELING

As mentioned in the "Motivations" section, our main objective is to develop a forecasting model for the total, gross earnings, i.e., "lordo totale, for each single day, i.e., starting from December 27th, 2022, to May 3rd, 2023, and to forecast said earnings for a future time period of seven days, for each one of the six restaurants. However, from the exploratory data analysis we discovered that the time series could actually be split in three different time periods, namely the pre-Covid-19 time period, the period during Covid-19 and more than two years after it.

For this reason, we thought it would not be appropriate to build a forecasting model on the whole time series data available, that is, from September 1st, 2018, to May 3rd, 2023, as forecasting is based on finding common regularities and systematics in the data according to parameters such as seasonality and trend, and each of the three time periods has its own distinct dynamics, characteristics and properties.

Therefore, we decided to build a forecasting model on a post-Covid-19 pandemic period, that is, from December 28th, 2020, to May 3rd, 2023, corresponding to approximately 2.4 years.

To achieve this task, we decided to build the forecasting model using the Prophet algorithm which, as mentioned in the previous section, was designed to provide accurate and intuitive forecasting of time series data. Moreover, the algorithm allows us to customize the model by providing additional domain-specific knowledge and incorporating external regressors that may influence the time series, e.g., weather, specific holidays and Covid-19 information.

PREPROCESSING

The dataset used for modeling now consists of 5136 observations, that is, the total, gross earning of the six restaurants recorded over time, and six variables containing information about: (1) date, (2) "lordo totale", (3) the restaurants, (4) weather, (5) the cities where the restaurants are located, and (6) lags for seven days.

Before actually developing the model and prepare the data, we decided to split the time series data of the restaurants into train and test sets as follows:

- train set, data from December 28th, 2020, to December 26th, 2022;
- test set, data from December 27th, 2022, to May 2nd, 2023.

An exception has been made for restaurant R002, as its first data was acquired ad the end of 2019 which, therefore, resulted in biased information given the imminent start of Covid-19 pandemic. For this reason, the data for this restaurant was split as follows:

- train set, data December 27th, 2021, to December 26th, 2022;
- test set, data from December 27th, 2022, to May 2nd, 2023.

After determining the time series cut points, we therefore proceeded to create different data frames for each one of the six restaurants, namely "R000", "R001", "R002", "R003", "R004" and "R005". Furthermore, for each one of them we split the data into train and test sets.

THE MODEL

Using a Prophet framework, we opted to develop a pipeline which enabled us to compute the prediction accuracy for plus one day forecast, that is, from December 27th, 2022, to May 3rd, 2023, obtained by iteratively increasing the train set. Following this pipeline, at each iteration of the training process, the algorithm moves forward by one day and, at the end of each iteration, it adds a current new day to the train set. In other words, at each subsequent iteration, the training dataset increases by one day.

We care to note that, before training, we developed the model so as to include a list containing Ramadan, additional holidays and Covid-19 information (which are stated in section "Exploratory Data Analysis & Enrichment").

MODEL EVALUATION

To acknowledge how the model developed actually performs, it is fundamental to reference an error metric. The metric we decided to use is the Mean Absolute Percentage Error (MAPE). MAPE is a common metric used to evaluate the accuracy of a forecasting model or predictive algorithm. It measures the average percentage difference between the actual and predicted values, and it is calculated for each individual observation and then averaged to obtain a single value that represents the overall accuracy of the model. It is expressed as a percentage, which makes it easier to interpret and compare across different datasets and models. MAPE is a relative measure of error and is particularly useful when dealing with data that has different scales or units. A lower MAPE indicates a more accurate model, while a higher MAPE suggests less accuracy.

Aiming to understand the accuracy of the model in forecasting the total, gross earnings, i.e., “lordo totale”, we calculated MAPE values for each one of the restaurants, which can be found in the following table.

	R000	*R001*	*R002*	*R003*	*R004*	*R005*
MAPE	17.74%	10.77%	9.41%	1.061725096534327e+20%	10.12%	9.05%

Table 2

Overall, we can see from Table 1 that MAPE scores are not too high. The MAPE score for restaurant R005, located in Voghera (PV), is 9.05%, and it suggests that the model is able to forecast the total, gross earnings, that is, “lordo totale” for one day, for said restaurant with an accuracy that is approximately around 90.95%. In the case of restaurant R001, located in Piacenza (PC), the model is able to forecast its “lordo totale” with an accuracy of 90.59%. Moreover, the accuracy of the model for the restaurants R004 (Stradella, PV) and R001 (Piacenza, PC), is of 89.88% and 89.23% respectively. While for restaurant R000, located in Montebello (PV), the MAPE score suggest that the model is able to forecast its “lordo totale” with an accuracy of 82.26%.

However, we can see that the accuracy of the model in forecasting the total, gross earnings of restaurant R003, located in Piacenza (PC) drops dramatically, as the score of the Mean Absolute Percentage Error equals 1.061725096534327e+20%, suggesting that, in this case, the model performs poorly. For this reason, we decided to not include this restaurant in the seven-day forecast (that will be introduced in the section), and it is our concern to point out that using another forecasting model could improve the score of the above-mentioned metric.

In order to understand how the forecast differs from the actual time series data, we performed graphical inspection, taking as reference the restaurant R001, located in Piacenza (PC), for simplicity.

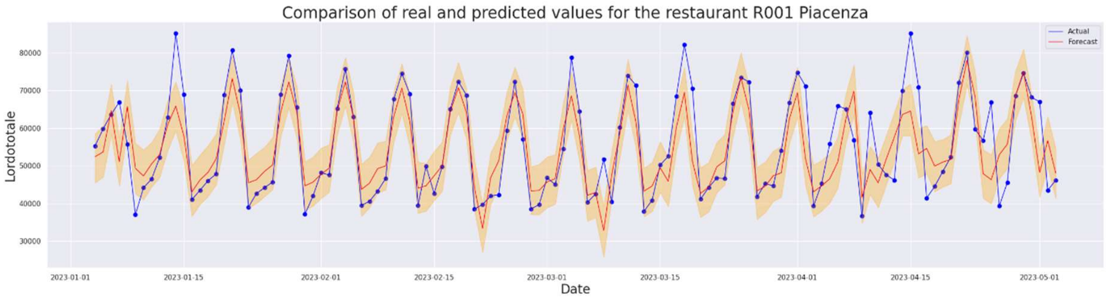
Comparison of real and predicted values for the restaurant R001 Piacenza

Figure 5

As we can see from Figure 5, the behavior of the forecast, for the time period that goes from December 27th, 2022, to May 3rd, 2023, does not seem to be significantly different from the actual time series data. Therefore, we can state that, also considering the MAPE score, the model built upon the Prophet algorithm appears to perform well.

SEVEN DAYS FORECAST

As already mentioned in section “Motivations”, the main goal of this project was to build a forecasting model that is able to predict the total, gross earnings of the restaurants. Keeping in mind that the model was trained on a time interval that goes from December 28th, 2020, to May 2nd, 2023, for the restaurants R000, R001, R004, R005, and on a time interval that goes from December 27th, 2021, to May 2nd, 2023, for restaurants R002, we predicted “lordo totale” for the time period that goes from May 4th, 2023, to May 10th, 2023.

The results of said prediction can be found in the following table.

Lordo Totale Predictions in €

Date	***R000 Montebello (PV)***	***R001 Piacenza (PC)***	***R002 Piacenza (PC)***	***R004 Stradella (PV)***	***R005 Voghera (PV)***
04/05/2023	28915.590	51430.712	43514.481	40216.751	42156.887
05/05/2023	38066.612	67139.502	56272.214	52457.718	54740.375
06/05/2023	50085.575	73640.604	66234.072	59648.363	51660.348
07/05/2023	47739.820	66738.166	67935.357	51134.058	45985.473
08/05/2033	35333.263	56713.757	52578.619	45374.573	42038.693
09/05/2023	27983.970	47760.884	51630.038	38187.700	39156.483
10/05/2023	29711.557	50692.499	44200.472	39516.821	41634.464

Table 3

To have a better understanding of how the total, gross earnings of the restaurants behave in these seven days, we decided to graphically inspect said forecast, which we display below.

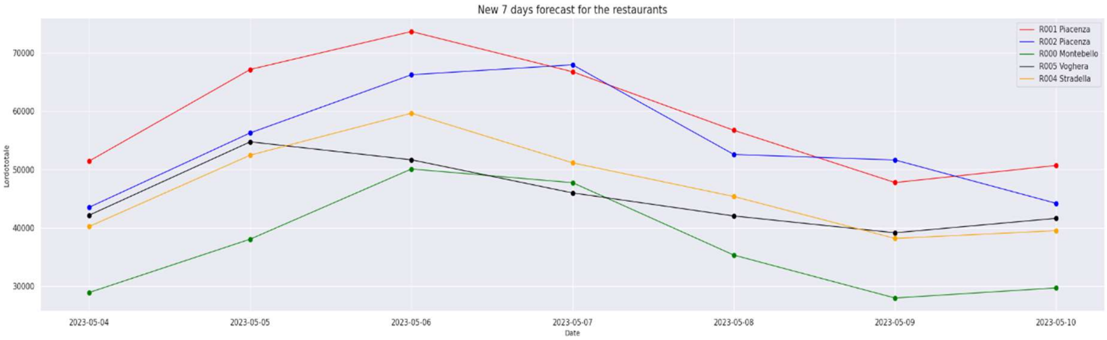

Figure 6

As we can see from Figure 6, “lordo totale” for all the restaurants appears to increase from May 4th, 2023, to successively reach its peak on May 6th, 2023, before it gradually decreases. This trend could be explained by simply looking at a calendar: as a matter of fact, May 6th, 2023, results to be a Saturday, which is more likely a day in which restaurants tend to earn more.

CONCLUSIONS

In this report we outlined and described the focal points of our analysis, which was carried out on a dataset containing daily records of the total gross earnings of six restaurants, on a time period of almost five years, going from September 1st, 2018, to May 3rd, 2023, with the aim of developing a forecasting model.

Throughout the exploratory data analysis, we were able to observe how the "lordo totale" of the restaurants change over time, and we discovered that external factors, that is, factors not present in the data originally available, could affect the behavior of the total gross earnings. This discovery led us to enrich the dataset with information pertaining to weather phenomena, Covid-19 pandemic and specific holidays.

Furthermore, we were able to observe that the "lordo totale" of each restaurant exhibits different dynamics and characteristics over three different time spans, that is, pre-Covid-19 period, the period during Covid-19, and the period after the pandemic; this led us to decide it would be appropriate to treat the three different time periods separately, and to develop the forecasting model only on the post-Covid-19 time period, i.e., from December 27th, 2022, to May 3rd, 2023.

By means of Prophet forecasting tool, we preprocessed the data and split them into train and test sets. We then proceeded to build a forecasting model which enabled us to compute the prediction accuracy for plus one day forecast, that is, from December 27th, 2022, to May 3rd, 2023, obtained by iteratively increasing the training set.

Additionally, we performed model evaluation by choosing as error metric the Mean Absolute Percentage Error (MAPE). The results obtained from this procedure led us to the conclusion that the model is able to accurately forecast the total gross earnings of the restaurants R000, R001, R002, R004, and R005, with the exception of the restaurant R003, which we then decided to exclude from the final step of our analysis.

Finally, we applied the forecasting model previously trained to forecast the "lordo totale" of the above-mentioned restaurants for a total of seven future days, that is, starting from May 4th, 2023, to May 10th, 2023. The results of this procedure led us to the conclusion that the restaurants are likely to earn more during the weekend than during weekdays.

REFERENCES

1. https://github.com/JuliaTsymbal/DataScienceLabProject (at this link all the materials used to develop the project can be found);

2. http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1#;

3. https://www.ilmeteo.it/portale/archivio-meteo.

Additional references:

a) https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775;

b) https://facebook.github.io/prophet/;

c) https://nbviewer.org/github/nicolasfauchereau/Auckland_Cycling/blob/master/notebooks/Auckland_cycling_and_weather.ipynb.