Data Science Lab Project: restaurants time series

Matteo Corona, Francesca Corvino

Abstract

This report presents a project for the "Data Science Lab" course conducted on the historical sales data of various restaurants. The main objective of this project was to apply data analysis techniques to understand the underlying patterns in restaurant revenue over time and identify any key factors influencing variations in gross revenue and daily receipt counts. In the initial phase of the project, the data related to the daily revenue of the studied restaurants was prepared. Subsequently, an initial exploratory analysis was conducted to identify trends, seasonality, and anomalies in the data. Graphs and descriptive statistics were used to obtain a preliminary view of the prominent features of the time series. In the modeling phase, various approaches were implemented for time series modeling and forecasting of future revenue. Forecasting models based on traditional methods such as ARIMA and SARIMA were used, as well as more complex models based on recurrent neural networks (RNNs). The analysis results revealed interesting seasonal patterns within restaurant revenue, with peaks during festive seasons and more stable trends during the rest of the year. The forecasting models demonstrated good performance in capturing these seasonal variations but faced challenges in predicting sudden variations due to exceptional events. In conclusion, this data science project provided a detailed overview of the restaurant revenue trends over time and demonstrated the effectiveness of various analysis and modeling techniques for time series forecasting.

1 Introduction

At the heart of every restaurant, there's a story to be told. Each tablecloth laid on a table, every smile exchanged among diners, and every dish crafted with care encapsulates a unique chapter of human experiences. Yet, there's another aspect of these stories that often remains hidden: the data. Behind every service, every order, and every receipt lies an invisible web of numbers and trends that can reveal much more than one can imagine. The goal of this project was precisely to delve into these data to analyze and characterize the financial journey of some restaurants, especially after a period of severe crisis in the restaurant industry due to the COVID-19 pandemic.

2 The dataset

The dataset contains information about six restaurants located in different cities[1]. The attributes contained in the dataset are:

- date

[1] The restaurant locations were added by integrating information from another dataset

- receipts
- gross total
- restaurant
- city

From the time series (as shown in Figure 1), some values were removed:

1. **First 8 months**: The values for the first 8 months were all null, except for the first day of each month. This was likely because initially, the values were recorded on a monthly frequency rather than daily. To maintain a constant sampling frequency, it was decided to keep only the daily data.
2. **Last 50 days**: The last 50 days (May and June 2023) were all null values. These are clearly outliers, representing days that were not recorded and do not reflect the normal trend of the restaurants.
3. **COVID-19 period**: The project's goal is to model the "normal" trend in gross revenue for the restaurants. For this reason, an exceptional period like the COVID-19 pandemic was not considered relevant and was removed as an outlier. Additionally, the presence of a block of null values in the series would complicate the analysis of stationarity and could negatively affect model performance. In conclusion, to remove the COVID-19-related period, it was decided to study the series starting from May 7, 2020 (the first date after the end of the lockdown period). This approach aims to characterize the period of financial growth and return to normalcy experienced after the pandemic [Data Imputation Demystified | Time Series Data].

3 Exploratory data analysisi

Before conducting more in-depth statistical analyses, it is essential to perform exploratory data analysis (EDA). A first step is to visualize the temporal trend of the series. Figure 1 displays the trend of "lordototale" over time for one of the six available restaurants.

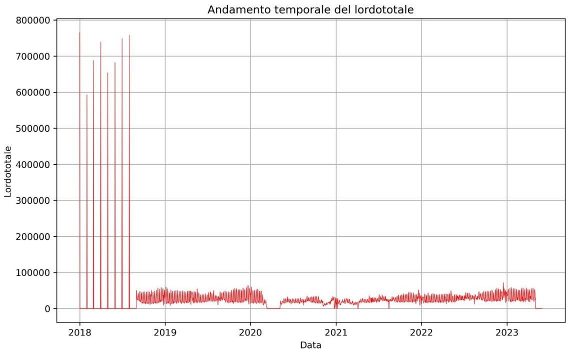

Figure 1: Time series for restaurant R000

From the graph in Figure 1, an obvious anomaly is immediately noticeable in the first half of the year 2020. This anomaly corresponds to the closure days due to the COVID-19 pandemic. As mentioned earlier, some values in the series were removed to facilitate analysis and avoid influencing model performance. The graph in Figure **??** displays the time series for restaurant R000 after this step. Regarding the variables available in the dataset, it is expected that the "lordototale" and "scontrini" variables would have a strong linear correlation. This was confirmed by the coefficient of determination (R^2), which is equal to 0.91. The scatterplot in Figure 2 illustrates the relationship between these two variables.

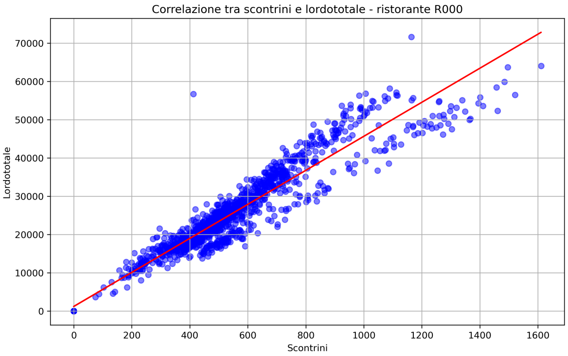

Figure 2: Correlation between lordototale and scontrini - restaurant R000

The slope of the interpolation line in Figure 2 is approximately 44.5. This value can be interpreted as the average price of a receipt in restaurant R000. However, despite the high R^2 value, indicating a strong positive correlation, there seems to be a division of data points along two different lines. This suggests that there might be discrepancies between data before and after COVID-19, between weekdays and weekends, or other factors. This aspect will be explored further.

Since the two variables are highly correlated, it is possible to focus the study on one of them. Considering the "lordototale" variable, it is useful to compare the six restaurants. This comparison was done through a density distribution plot of "lordototale" for each restaurant (Figure 3) and a boxplot that compares the "lordototale" variable for each restaurant (Figure 4).

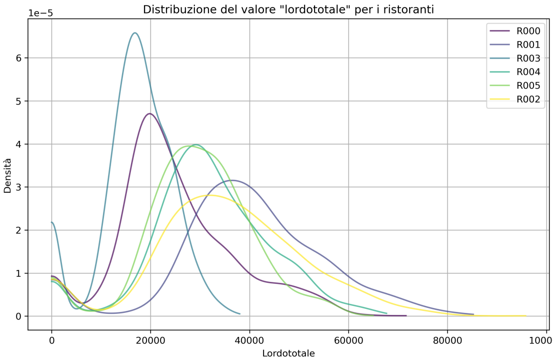

Figure 3: Density distribution of lordototale

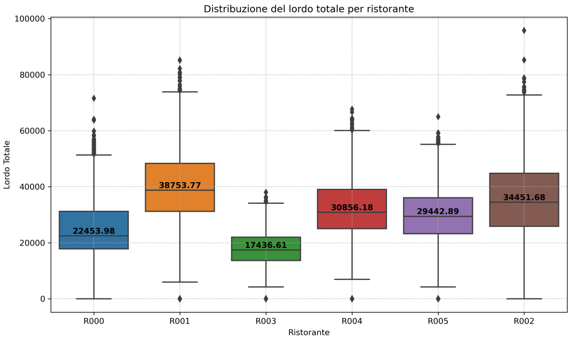

Figure 4: Boxplot of lordototale

From the graphs in Figures 3 and 4, one can appreciate the data's variability: for instance, restaurant R001 has a higher median "lordototale" but is also among those with lower density, indicating it could be a very expensive restaurant. At this point, as highlighted earlier, it is useful to compare "lordototale" concerning different types of days (weekends, weekdays, and holidays). Figure 5 shows the difference between these three groups.

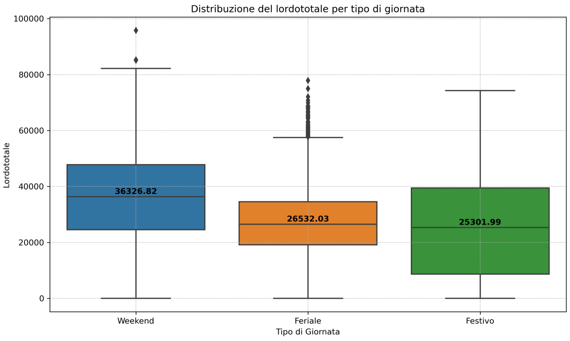

Figure 5: Boxplot of lordototale - days

By conducting an ANOVA test, it is found that the F-statistic value is 189, while the p-value is on the order of 10^{-81}. Since the p-value is practically 0, it can be concluded that there are significant differences in at least one of the analyzed groups. In other words, there is strong statistical evidence to assert that the group means are not all equal. The very high F-statistic supports this conclusion, indicating that the variation between groups is significantly larger than the variation within groups. It can be observed that "lordototale" seems to be higher on weekends.

4 Stationarity analysis

To analyze a time series and apply common models like ARIMA, it is necessary to verify if the series under analysis satisfies the stationarity conditions. First, it is useful to decompose the time series in analysis: the *decompose()* command in *R* allows you to split a time series into three main components:

- **Trend**: This component represents the general direction of the data over time. In this case, from the graph in Figure 6, it is evident that there is an increasing trend. This trend aligns with expectations: after the COVID-19 crisis period, there was a gradual financial recovery. The presence of an increasing trend suggests that a non-seasonal differencing term will be needed to make the series trend-stationary.
- **Seasonality**: This component represents cyclic or seasonal variations in the data. From the *seasonal* component of the graph in Figure 6, one can observe, for example, the presence of some peaks that repeat every year. In general, there is a repeating pattern, indicating that the series indeed has a seasonal component (suggesting that a SARIMA model accounting for seasonality might be beneficial).
- **Residuals**: This component represents the residual error that cannot be explained by the trend or seasonality. It contains random variations and fluctuations in the data. Ideally, the residuals should represent white noise: if they are not normally distributed around zero, it could affect the model's performance.

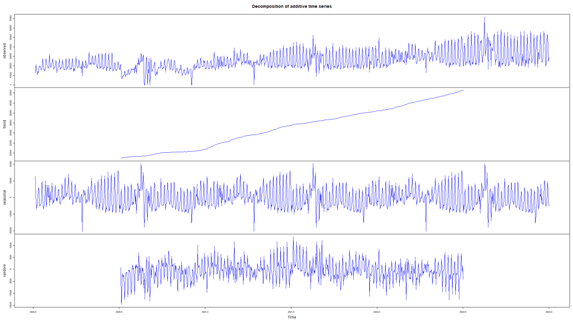

Figure 6: Time Series Decomposition - Restaurant R000

Following the decomposition analysis, two methods have been implemented to test the stationarity of the series:

- **ADF Test** (Augmented Dickey-Fuller test): This test starts with a null hypothesis stating that the time series is non-stationary. The test assumes that the time series has a unit root, meaning it exhibits a trend or structure that changes over time. The alternative hypothesis is that the series is indeed stationary. The result of the ADF test is a test statistic and an associated p-value. If the p-value is lower than a certain level of significance (e.g., 0.05), you can reject the null hypothesis and conclude that the time series is stationary.
- **KPSS Test** (Kwiatkowski-Phillips-Schmidt-Shin test): This test checks whether a time series is stationary around a constant or a deterministic trend. In other words, the KPSS test assumes that the series is stationary or exhibits random fluctuations around a

fixed level over time. The result of the KPSS test is a test statistic and an associated p-value. Unlike the ADF test, in the KPSS test, the goal is not to reject the null hypothesis. If the p-value is greater than a certain level of significance, you can conclude that the time series is stationary.

The test results are presented in Table 1. The ADF test indicates a stationary series, while the KPSS test indicates a non-stationary series. The two tests are in disagreement: this could imply the presence of a complex series [Statistical Tests to Check Stationarity in Time Series].

QUANTITY	ADF	KPSS
Statistic	-5.5	10.2
p-value	<0.1	<0.1

Table 1: Stationarity Tests

The analysis proceeds by formulating both an ARIMA model on the residuals and a SARIMA model on the original series to determine which one fits the data better.

5 The ARIMA model

In the context of the project under examination, a time series containing daily data collected over a four-year period was analyzed. The series exhibited a growing trend and significant seasonal patterns, manifested in three distinct cycles within the dataset. Given the complexity of the data, an ARIMA model was chosen as the starting point for time series analysis. This decision was influenced by the well-established reputation of the ARIMA model, known for its ability to capture both trends and other dynamic elements present in the data. To determine the optimal model parameters, the *auto.arima* algorithm was used, suggesting an ARIMA(5,1,0) model. Subsequently, a Walk-Forward validation technique was applied to assess the model's performance in predicting new data points. The Walk-Forward technique involved training the model on an initial portion of the dataset and then testing it on a future data point, repeating the process iteratively. While this method is computationally intensive, it provides an accurate estimate of the model's predictive capabilities. Despite these considerations, the ARIMA(5,1,0) model showed some limitations in reference to the metrics used, which will be analyzed later. The graph in Figure 7 displays the predictions of the total revenue obtained with the ARIMA model compared to the actual values of the time series.

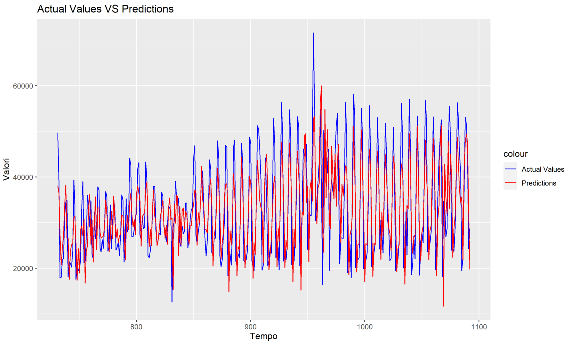

Figure 7: ARIMA Model Predictions - Restaurant R000

Following the ARIMA analysis, it was planned to explore more advanced models such as SARIMA, known for its effectiveness in handling situations of greater complexity, such as seasonality and structural changes.

6 The SARIMA model

In the case of the SARIMA model, a seasonal component with parameters (0,1,0) was introduced to capture the observed seasonal dynamics in the data. The choice of these parameters was guided by the need to make the series stationary in mean and stabilize the variance in the seasonal pattern. The choice of zero for both the seasonal autoregressive and seasonal moving average terms was deliberate to maintain a simple yet effective model, avoiding the risk of overfitting. This way, a compromise was sought between capturing the intrinsic seasonality in the data and maintaining the simplicity and interpretability of the model.

Similar to the ARIMA model, a *Walk-Forward* validation process was also employed for SARIMA. This choice was driven by the need to obtain an accurate assessment of the model, considering that the data exhibit both a growing trend and complex seasonality. The introduction of the seasonal component aimed to address some of the limitations highlighted in the ARIMA model, especially in terms of capturing seasonality. The effectiveness of this strategy will be subject to analysis for comparison with the ARIMA model to determine which of the two models provides the most accurate and reliable forecasts.

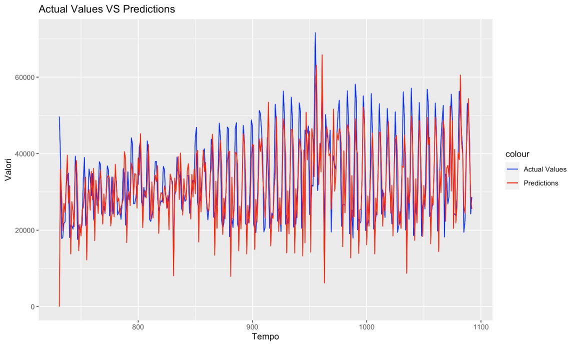

Figure 8: SARIMA Model Predictions - Restaurant R000

The graph in Figure 8 displays the predictions obtained with the SARIMA model for the most recent available data compared to the actual values of the series.

7 RNN

In the analysis process, after exploring ARIMA and SARIMA models, a Deep Learning model was also implemented. ARIMA and SARIMA models are traditionally used for time series analysis but may have limitations when it comes to capturing complex patterns or when data exhibits a nonlinear structure. Therefore, we adopted an approach based on Recurrent Neural Networks (*RNNs*). This allowed us to assess whether a Deep Learning model, specifically one based on *RNNs*, could provide more accurate forecasts compared to ARIMA and SARIMA models. Overall, the use of an *RNN* enriched the analysis, enabling us to further explore the potential of Deep Learning models for time series forecasting. The implemented model is quite simple and consists of a single *SimpleRNN* layer. It uses the *ReLU* activation function, its loss function is *mean squared error*, and it is configured with *Early Stopping* to prevent overfitting.

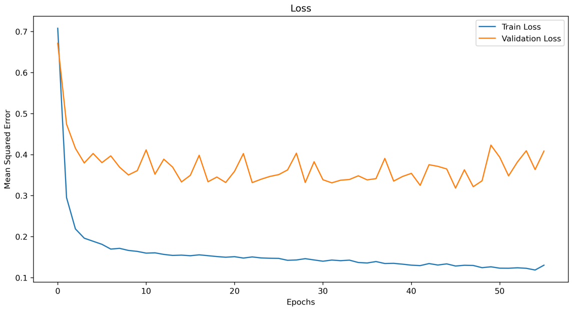

Figure 9: Loss function in RNN model

The graph in Figure 9 shows the loss function for the implemented *RNN* model. Notice that there is a discrepancy between the training loss and the validation loss. This discrepancy represents the minimum that could be achieved with this model. Even when attempting to increase the complexity of the model architecture or introduce additional layers like Dropout layers, this gap did not decrease. On the other hand, the graph in Figure 10 provides a visual comparison between the predictions and actual values for restaurant R000. At first visual analysis, this model appears to yield very satisfactory results, but these expectations will be verified by evaluating the metrics.

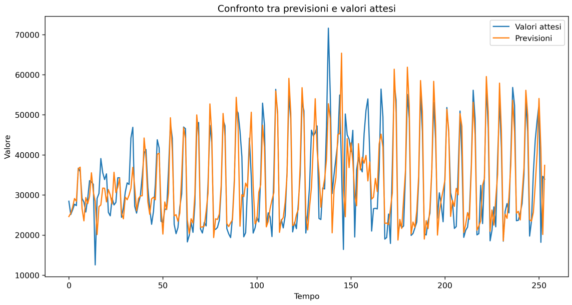

Figure 10: Predictions for restaurant R000 with RNN

8 Model comparison

The analysis of results and evaluation metrics, including RMSE, MAE, MPE, and MAPE, has provided a comprehensive overview of the performance of each model. This comparison has helped determine which model was able to produce more accurate forecasts, consequently aiding in making informed decisions for time series analysis. In Table 2, the metrics for each model are presented. It can be observed that the RNN model appears to be the best performer across all considered metrics [Forecasting Methods Showdown: ARIMA vs. RNNs].

model	RMSE	MAE	MPE	MAPE
ARIMA	7155	5534	-4.68	18.75
SARIMA	8644	6461	-3.96	22.76
RNN	6195	4316	-3.25	14.15

Table 2: Model metrics

9 Predictions

After comparing the models and determining the most suitable model, forecasts were made for all 6 restaurants using the RNN model. The obtained forecasts are shown in Figure 11.

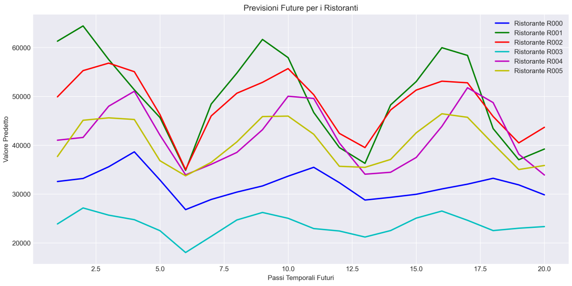

Figure 11: Total revenue forecasts for all restaurants (30 days)

The values in Figure 11 refer to forecasts for the 20 days following the last date in the dataset. The total revenue for the 6 restaurants in the considered period was also calculated, and the results are reported in Table 3.

RESTAURANT	TOTAL REVENUE [€]
R000	638,076.75
R001	999,682.81
R003	970,180.50
R004	473,346.25
R005	837,878.44
R006	809,507.50

Table 3: Cumulative future forecasts

10 Conclusions

During the analysis of time series data related to the total gross revenue of the restaurants, it became evident that the ARIMA and SARIMA models provided less than satisfactory performance. This outcome is largely attributed to the complexity of the series, characterized by a rising trend and strong seasonality, making it challenging for these models to handle. On the other hand, the approach based on Recurrent Neural Networks (RNNs), using models like SimpleRNN, proved to be more adept at capturing the dynamics of non-stationary series and managing the seasonal component. Consequently, the RNN models delivered more accurate forecasts. In conclusion, the RNN model emerged as the preferable choice for forecasting the examined time series, given the complexity of the data. It is important to emphasize that the effectiveness of a model always depends on the nature of the data, and in this case, the adaptation of RNNs to the dynamics of non-stationary series yielded better results compared to the ARIMA and SARIMA approaches.