E-Commerce sales forecasting with interactive visualisation of the predictions

Davide Ghilardi, Fabio Salerno

Abstract

Accurate sales forecasting is crucial for the success of ecommerce businesses, offering several advantages such as effective resource allocation, demand forecasting, budgeting, inventory management, and targeted marketing. This project aims to forecast future sales for an ecommerce store using various techniques and identify the best-performing predictive model. Additionally, an interactive dashboard will be designed to present the forecasting results in a clear and accessible manner. The project involves pre-processing the data, testing different models on various time spans, evaluating the results using performance metrics, and selecting the optimal predictive model for deployment. The developed dashboard will provide the ecommerce managers with valuable insights to improve their strategic decision-making based on the sales forecast.

Keywords: E-commerce; Time series; Machine Learning

Contents

1. Introduction

The ability to accurately forecast sales for an ecommerce business is crucial and offers several advantages such as:

- **Resources allocation**: sales forecasting enables e-commerce businesses to plan their operations effectively.
- **Demand forecasting**: providing insights into customer demand patterns, allowing businesses to anticipate trends and adjust their product offerings accordingly. It helps identify popular products and categories, allowing for targeted marketing and product development strategies.
- **Budgeting**: allows businesses to estimate revenue and project cash flow, facilitating the development of realistic budgets. With a clear understanding of sales expectations, businesses can make informed decisions regarding investments, marketing budgets, and expense allocations.
- **Inventory management**: optimize inventory levels by providing insights into anticipated demand for different products and categories. This minimizes the risk of overstocking or understocking, reduces carrying costs, and improves overall inventory turnover.
- **Marketing**: by knowing when sales are expected to peak, businesses can strategically time their marketing efforts to maximize the impact. It helps in optimizing promotional activities, such as offering discounts, running targeted advertisements, or launching new products during high-demand periods.

So accurate sales forecasting can positively contribute to the success of a business, by exploiting data-driven decisions enhancing efficiency, profitability and customer satisfaction.
The aim of the project is to forecast the future sales of an ecommerce store through different techniques, identifying the best performing predictive model for each time frame to be used by the ecommerce managers to improve their strategic business decisions. Furthermore, an interactive dashboard will be developed to present the forecasting results in a clear and accessible manner.
First of all, the data were pre-processed, so as to make it effective and optimal for the purposes of the project; predictive models, indeed, require input data modeled in a standard manner. Once the data were processed, the focus shifted to testing the various models, on different time spans (quarterly, monthly and weekly) then the results obtained were evaluated through MAPE (mean absolute percentage error). Finally the best predictive model was chosen to deploy the sales forecast; then it was designed a dashboard for presenting to the managers the results obtained.

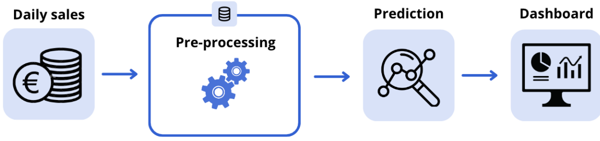

Figure 1 Schema of the project workflow.

2. Data

The dataset was given in CSV format and contains the daily sales of the ecommerce for each sector, in particular there are 26615 observations, the columns are:

- **Date**: which contains the date of each observation, in the format: Day/Month/Year.
- **Total**: amount in euro of the revenue coming from a specific sector on that day.
- **Sector**: identifying the specific sector of the ecommerce.

For many sectors, the number of observations is limited, thus to have consistent and reliable results it was decided to consider the three sectors with the highest number of observations, which are: fishing, casual and football.

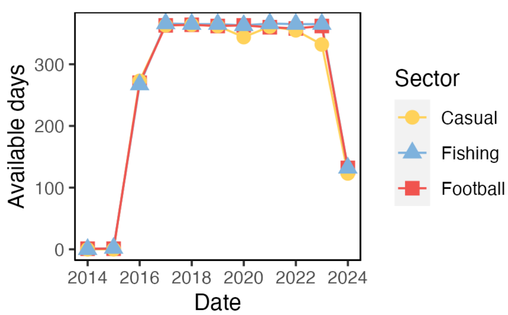

Figure 2 Number of observations for the three sectors through the years.

On the x-axis we have the date, on the y-axis the number of observations. So for each sector taken into account (fishing, casual and football) this chart represents the total number of observations grouped by year. In particular, for 2014 the number of observations is zero, while from the following year this figure grew for both of the three sectors. In the time range between 2016 and 2022 the information available is at its peak. For 2023, since the year is not yet over, the amount of information available only covers the months from January to May.
Taking the chart and the resulting considerations into account for analysis purposes, it was decided to look at the data in the time frame between 2016 and 2022.

A. Pre-processing

The available data were not particularly problematic except for the presence of missing values for some days of the year.

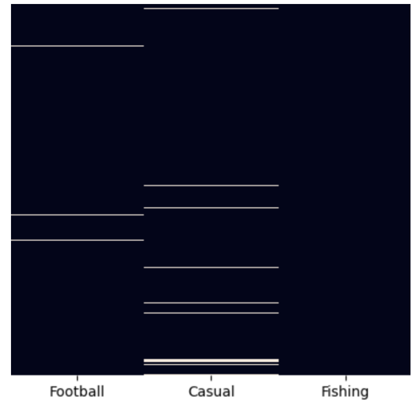

Figure 3 Pixel map for the missing values of each sector's time series.

This visualization figures a pixel map where on the y-axis there's the time period taken into account (2016-2022) and on the x-axis there are the three sectors analysed (fishing, casual and football). This map indicates for each sector if the observation is missing in a specific day of the time range considered. The fishing sector doesn't present any missing value, the football sector presents just three observations missing, while the casual sector is the one more problematic in terms of missing values.
This figure gives us an overview of the status of missing data in each sector; several solutions were examined to solve this issue, including: considering the days with missing data as days in which the revenues for that specific sector was zero; another possibility was to consider those days with missing data days in which, due to some error in the system, it was not possible to save the record of the earnings.
Since the missing data doesn't show a systematic trend even inter-sectoral, and since missing values aren't associated to particular days of the year, it was decided to give more consideration to the hypothesis that sales were not recorded, and for these reasons it was decided not to consider the receipts of those days as zero, but instead to estimate the value of the missing observations by interpolation.

3. EDA

The next step was dedicated to the creation of the datasets that will then be used in the respective models. Once the NA values were treated, the data was splitted for each sector taken into account (fishing, casual and football); then, the data was aggregated considering different time granularity, in particular the datasets were created for weekly, monthly and quarterly sales.
To get a better understanding of the sales data for each sector, an exploratory analysis was carried out on all the time series. In this section, the most relevant results are shown.
Assessing changes in trend and variability in a time series is important for understanding the underlying patterns and dynamics of the data. A preliminary visual inspection can be done to look after shifts, patterns, or irregularities in the data over time; changes in the slope, amplitude, or frequency of the data points may indicate variations in trend or variability.

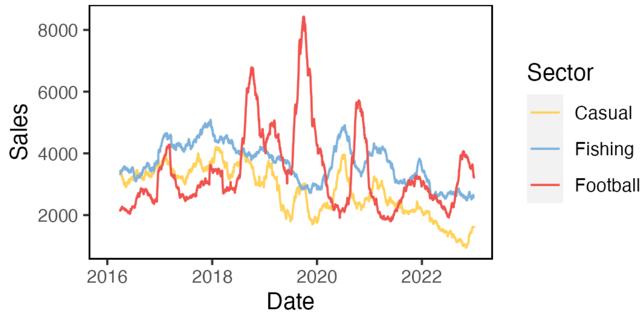

Figure 4 90 day moving average of the three sector's time series.

The graph displays for each sector taken into account its time series along the time range considered (2016-2022). On the x-axis it's displayed the time while on the y-axis there are the revenues. The time series is a 90 day moving average, which acts as a low pass filter to reduce short-term (high-frequency) fluctuations and better visualize the overall pattern of the data.
Overall all the three sectors present a trend which is initially positive but from 2019 turns negative. Furthermore the football sector appears to be affected by a strong seasonality effect, which is also present in the other two sectors.

A. Seasonality

Seasonality refers to systematic fluctuations that occur periodically within the 12 months of the year. Seasonality in time series can arise due to a variety of factors, including natural phenomena, human behavior, and calendar effects. For example, retail sales may exhibit seasonality with higher demand during holidays. Identifying and understanding seasonality in time series data is important for accurate forecasting and predictions. To identify patterns and variations within each season or time period and to gain a deeper understanding of the underlying seasonal behavior of each sector it was used the following plots:

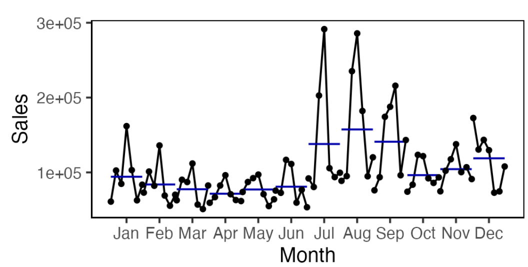

Figure 5 Seasonal sub-series plot of Football monthly sales.

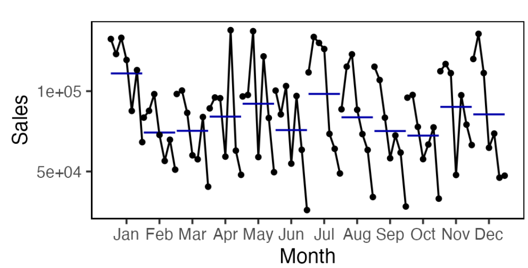

Figure 6 Seasonal sub-series plot of Casual monthly sales.

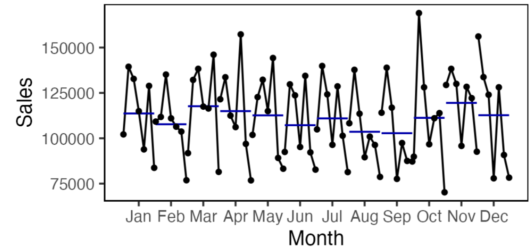

Figure 7 Seasonal sub-series plot of Fishing monthly sales.

The seasonal subseries plot is a graphical tool used to visualize and analyze seasonality. The time series is divided into monthly subsets based on the seasonal period on the x-axis and the y-axis representing the revenues; so for each month we have a subset time series of each specific month through the years. It helps visualize the magnitude and patterns of seasonality within each month, making it easier to observe any recurring patterns, trends, or anomalies.
Overall all the sectors are affected by seasonality, in particular the football sector is the one more affected by this phenomena, from the plot it stands the difference in earnings during the summer compared to the other months, this result may be due the warmer weather which incentive customers purchasing football items.

B. Correlation

The linear correlation coefficient measures the strength and direction of the linear relationship between two variables. The value of r ranges between -1 and 1. A positive value indicates a positive linear relationship (as one variable increases, the other tends to increase), while a negative value indicates a negative linear relationship (as one variable increases, the other tends to decrease). The closer ρ is to -1 or 1, the stronger the linear relationship. A value of 0 indicates no linear relationship between the variables.

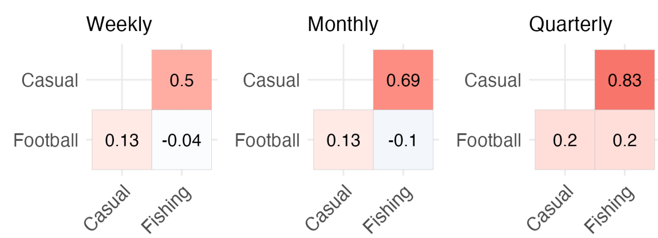

Figure 8 Correlations between sectors sales time series.

The above chart includes several correlation plots, also known as correlation matrices, one for each time frame. They visually displays the correlation coefficients between the three sectors in terms of revenue. Each variable is represented by a row and a column, and the cells in the plot contain the correlation coefficients between the corresponding pairs of sectors. The correlation coefficients are color-coded with a divergent colormap that ranges from blue ($\rho = -1$) to red ($\rho = 1$) passing from white ($\rho = 0$).
Focusing on quarterly sales, it can be noticed that the fishing and the casual sectors are highly positively correlated with each other

(0.83), which means that when there's an increase/decrease in the sales of one sector, the other tends to behave in the same way in terms of sales results. The other combinations of sectors are almost uncorrelated with a positive correlation coefficient equal to 0.2.

Due to these results it was done an in-depth study was done about the correlation between sectors separately considering each quarter, aiming to understand how the relationship between sectors changes through the quarters of the year.

The casual sector and the fishing sector remain highly positively correlated along the quarters. Considering the linear relationship between the remaining sectors (football-casual, football-fishing) for the first quarter they are both highly positively correlated; while during the remaining quarters they are almost uncorrelated. In conclusion the revenues of the three sectors in the initial part of the year are positively linearly correlated between each other, and through the year the level of linear relationship tends to decline except between the casual and the fishing sector.

C. Correlograms

Correlograms are useful plots to evaluate seasonality at different levels of frequency. The ACF (Autocorrelation Function) measures seasonality at a given frequency as the correlation between the time-series and its lagged version corresponding to that frequency (e.g. we expect a monthly time series with annual seasonality to show high correlation at lag 12). The ACF plot is often used in conjunction with the PACF (Partial Autocorrelation Function) which measures the correlation between a time series and its lagged values while controlling for the intermediate lags, excluding their effects on the correlation values.

Before computing ACF and PACF the time series were differentiated to make them stationary.

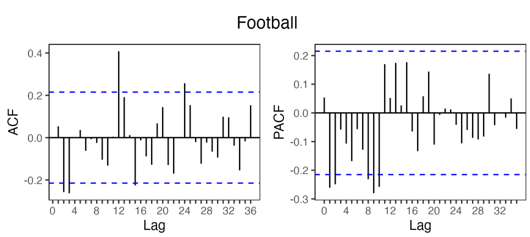

Figure 9 Correlograms of Football monthly sales.

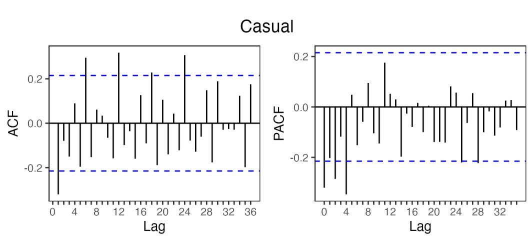

Figure 10 Correlograms of Casual monthly sales.

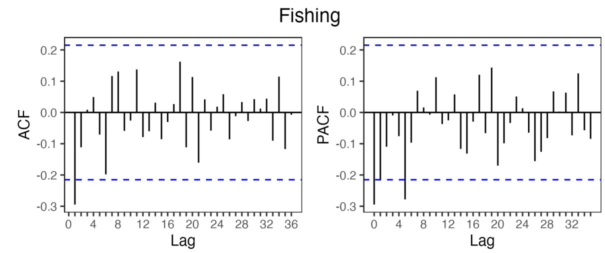

Figure 11 Correlograms of Fishing monthly sales.

The ACF and the PACF plots display respectively the autocorrelation coefficients and the partial autocorrelation coefficients (y-axis) against different lag values (x-axis). A significant positive or negative autocorrelation value at a given lag suggests that the time series is correlated with its past values at that lag. The blue dashed lines in the ACF and PACF plots represent confidence interval and only lags associated with values that overcome them can be interpreted as significantly different from zero; viceversa, anything within the blue lines is statistically not different from zero.

Taking into account the correlograms of the monthly time series we can say that:

- for the football sector, there are several autocorrelations that are significantly non-zero, indicating the presence of seasonality. In particular the ACF shows positive values at the time lags 12 and 24, so presumably the series is associated with annual seasonality;
- for the casual sector there's a high degree of autocorrelation for lag 1, 6, 12, 18 and 24, indicating signs of semestral seasonality; the PACF shows significant values until lag 4;
- the ACF of the fishing sector indicates high autocorrelation only for lag 1, while in the PACF there are significantly non-zero values at lag 1 and lag 5.

The results of these analyses on the time series were exploited to gain a deeper understanding of the data and also to identify significant patterns to better fit the predictive models.

4. Methods

As mentioned above, for each sector (fishing, casual and football) forecasts and assessments were made with the different time aggregations (weekly, monthly and quarterly). In particular the following models were developed:

- SARIMA
- XGBoost
- LSTM

A. Train Test split

Since to train and evaluate time series the ordering of the data is fundamental, we couldn't opt for a K-fold CV, and the decision of the dataset split was forced to a classical train test split.

The training set included 80% of observations, which means:

- From 2016-01-03 to 2021-08-01 for weekly data.
- From Jan 2016 to July 2021 for monthly data.
- From Q1 2016 to Q3 2021 for quarterly data.

The test set included contained the following observations:

- From 2021-08-08 to 2022-12-31 for weekly data.
- From Aug 2021 to Dec 2022 for monthly data.

- From Q4 2021 to Q4 2022 for quarterly data .

While for the SARIMAX model, only a single time series, plus eventual exogenous variables, is required to fit the model, for the other two, XGBoost and LSTM, the dataset had to follow an additional preparation step. In particular, the dataset had to be transformed in the classical classification/regression format with X and y, where the first is a matrix of shape $m \times d$, where m is the number of observations and d the number of covariates, and y is a vector of length n. The X matrix is created by shifting the observations in time according to two parameters: n and k; n defines the maximum lag to be included in the forecast, while k indicates the number of lags that has to be skipped from one lag to the other. The chosen values of the two parameters were $n = 52$, $k = 6$ for weekly models, $n = 12$, $k = 2$ for monthly models, and $n = 4, k = 1$ for quarterly models.
When using time series in real world problems, it's important to decide what strategy has to be followed to handle new data obtained from the business process. In fact, while new information is continuously produced and can be directly fed to the models, this requires a lot of computational power because a complete fit has to be ran at each timestep. Our solution to this problem was to set a parameter to control the duration of the validity of each model, after which a new fit on all the new observations was required. The values of the parameter were 8 for the weekly models and 1 for the monthly and quarterly models.

B. SARIMAX

It stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors and refers to a family of models that apply a stochastic approach to stationary time-series with a seasonal component. It's indicated with

$$\text{SARIMAX}(p, d, q)(P, D, Q)^s$$

and defined by seven hyper-parameters:

- p is the order of the autoregressive component;
- d is the number of differentiations;
- q is the order of the moving average component;
- P is the order of the seasonal autoregressive component;
- D is the number of seasonal differentiations;
- Q is theorder of the seasonal moving average component;
- s is the seasonal order.

The SARIMAX model can be divided into a classical and a seasonal part, both of which include an AR and a MA component. The first models past values of the time-series, while the second deals with the residuals of the regression. Lastly, differentiation has to be managed to make the time series stationary. The equation of a SARIMAX$(p, d, q)(P, D, Q)^s$ model can be written as

$$(1 - \boldsymbol{\phi}^T \boldsymbol{L}_p)(1 - \boldsymbol{\Phi}^T \boldsymbol{L}_P)\Delta^d \Delta_s^D y_t = (1 - \boldsymbol{\theta}^T \boldsymbol{L}_q)(1 - \boldsymbol{\Theta}^T \boldsymbol{L}_Q)\epsilon_t$$

where:

- $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ are the AR coefficients;
- $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_P)$ are the seasonal AR coefficients;
- $\boldsymbol{L}_p, \boldsymbol{L}_q, \boldsymbol{L}_P, \boldsymbol{L}_Q$ are the lag operator of each component (e.g. $\boldsymbol{L}_p = (L, L^2, \ldots, L^p)$, where L is the single lag operator)
- Δ is the differencing operator;
- Δ_s represents the order of seasonal differencing;
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ are the MA coefficients;
- $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_Q)$ are the seasonal MA;
- y_t represents the time series at time t;
- ε_t represents the error term at time t.

The X of SARIMAX comes from the addition of exogenous variables that try to explain the coefficients of the model. In particular, we also gave the model information about the month and the quarter. To build a SARIMA model, two main approaches can be followed: the first relies on the inspection of correlograms and the application of several heuristics to choose the right parameters; the second approach is based on information criteria such as AIC, and tries to optimize its value.
Since our models weren't too computational expansive, we decided to take the second approach and build a model for each combination of sector and time frame. The parameters were optimized with a stepwise search by minimizing the AIC. The obtained models' parameters are resumed in the following table:

Table 1 SARIMAX models

	SARIMAX models		
	Football	Casual	Fishing
W	$(3,1,1)(1,0,0)^{52}$	$(0,1,1)(0,0,1)^{52}$	$(3,1,2)(2,0,0)^{52}$
M	$(1,0,0)(0,0,0)^{12}$	$(1,1,1)(0,0,1)^{12}$	$(0,1,2)(0,0,0)^{12}$
Q	$(0,0,0)(1,0,0)^{4}$	$(0,1,0)(0,0,0)^{4}$	$(0,0,0)(1,0,0)^{4}$

C. XGBoost

XGBoost [1] is a machine learning algorithm which applies gradient boosting [2] to decision trees. Gradient boosting is a powerful technique that train weak models iteratively, and each subsequent model is built to correct the mistakes made by the previous ones. In this way, the overall model can focus on observation that are difficult to predict. Moreover, XGBoost provides insights into feature importance by ranking the importance of input features based on how much they contribute to the predictive performance of the model. Lastly, it includes L1 and L2 regularization techniques to prevent overfitting and improve generalization by controlling the complexity of the model.
We decided to choose default configuration of XGBoost regressor which is defined with the following specifications:

- Number of decision trees: 100
- Max depth of each tree: 6
- Weight of the L2 regularization (lambda): 1
- Weight of the L1 regularization (alpha): 0.

D. LSTM

Long Short-Term Memory (LSTM) [3] are a sophisticated type of recurrent neural network (RNN) [5] specifically designed to handle temporal data. It belongs to the group of deep learning techniques, which focuses on building and training neural networks with multiple layers. These networks are designed to automatically learn hierarchical representations of data, allowing them to handle complex problems. The LSTM architecture handle time series efficiently by deciding at each training step which information has to be preserved and what has to be thrown away. Our network had the following structure:

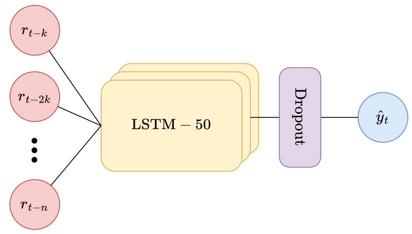

Figure 12 The architecture of the LSTM model.

We decided to implement a network with a single layer of 50 LSTM with a 20% dropout [6], followed by a single dense layer to make the prediction. We compiled it with Adam [4] as optimizer, and MSE as loss.

E. Evaluation (MAPE)

In assessing and selecting the most suitable and appropriate and precise for the purposes of the project, it was decided to use MAPE as the evaluation metric for the models.
The Mean Absolute Percentage Error (MAPE), is a measure of prediction accuracy of a forecasting method in statistics.

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_i}{y_t}\right| \times 100\%$$

Where T is the length of the time series, y_t is the actual value, and $\hat{y}_i$ is the forecast value. Their difference is divided by the actual value y_t. The absolute value of this ratio is summed for every forecasted point in time and divided by the number of fitted points T. MAPE has a range between 0 to infinity and provides a measure of the relative forecasting error in percentage terms. It gives an indication of the average magnitude of the errors compared to the actual values. A low MAPE (under 20%) indicates a good forecasting accuracy, while a higher MAPE (over 50%) suggests a larger average deviation between the predicted and actual values. For instance, it is sensitive to extreme values or outliers in the data and can become undefined if the actual values are zero or close to zero.

5. Results

For each sector (fishing, casual and football) forecasts and assessments were made with the different time aggregations (weekly, monthly and quarterly) and the following results were obtained.

Table 2 MAPE weekly data

	MAPE		
	Football	Casual	Fishing
SARIMAX	**39.34%**	**42.96%**	**22.00%**
XGBOOST	50.37%	72.50%	37.31%
LSTM	51.15%	87.66%	36.1%

Table 3 MAPE monthly data

	MAPE		
	Football	Casual	Fishing
SARIMAX	35.45%	**26.08%**	**11.34%**
XGBOOST	33.68%	42.20%	21.40%
LSTM	**24.85%**	44.42%	25.28%

Table 4 MAPE quarterly data

	MAPE		
	Football	Casual	Fishing
SARIMAX	24.09%	**16.22%**	28.41%
XGBOOST	**14.02%**	19.56%	**23.07%**
LSTM	18.15%	47.76%	36.67%

Considering the weekly time series for every category, the SARIMAX achieves the best results in terms of MAPE, with an average error between real and forecasted data of 39.3% for the football sector, 43.0% for the casual, and 22.0% for the fishing sector. Weekly forecast have a MAPE all above 20% which isn't good, and a possible explanation of this could be the fact that we decided to choose a period of 8 weeks before refitting the model on new data. For the monthly time series, the LSTM achieves the best results for the football sector (24.9%), while the SARIMAX obtained it for the casual and the fishing sectors (26.1% and 11.3%). For the quarterly time series the SARIMAX achieves the best results for the casual sector (16.2%), while the XGBoost shows the best performances for the football and the fishing sectors (14.0% and 23.0%).
Overall, considering the results obtained, the model that on average has the worst performance is the LSTM. On the other hand, the model that best predicts future revenues for each sector is the SARIMAX which is able to achieve on average a forecasting error lower than the others. The following plot shows the real and forecasted values for the SARIMAX model for the monthly time series:

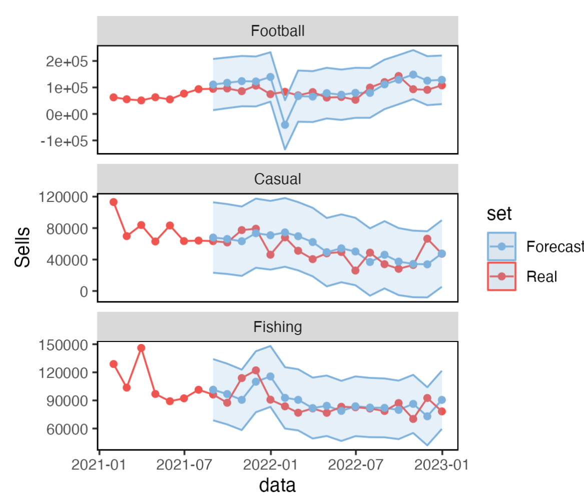

Figure 13 SARIMAX time series with forecast and CI.

In the chart, the red line with dots represents the real time series, while the blue one indicates forecasted values. The blue area is the 95% confidence interval of the forecasting, which indicates the range of plausible values of the revenue predicted by the model. It can be noticed that the confidence interval is very large with respect to the range of values so the forecast is highly volatile (notice that for the casual sector the CI includes negative values).

6. Dashboard and conclusions

The project aim was to forecast the future sales of an ecommerce store through different techniques and different time frames, and we found the SARIMAX to be the one that has the best overall performance. We are satisified by the obtained results, in particular by the monthly and quarterly predictions which achieve an average percentage error almost not below 25%.
Another objective of the project was to give insights to the ecommerce managers and improve their strategic business decisions. In respect of that, an interactive dashboard was designed for presenting to the managers the results obtained. The dashboard is a mockup of a possible dashboard used for presenting in an intuitive and refined way the results of the forecasts, aiming to assist the managers of the ecommerce on their future decision strategies by giving a refined overview of the results. The dashboard designed can be accessed by the following **LINK** and it is divided in the following sections:

- Next time period forecast;
- total sales trend;
- revenue forecast by sector;
- top three sectors monthly forecast.

In the first section, which is located on the upper part of the dashboard, there are three boxes: the first one is related to the current monthly sales and the next month's forecast compared with the total sales of the previous year in the same period. Is possible to switch between the current month and the next month prediction through a button located on the top right of the box. The same way of reasoning is applied to the quarter sales box, which indicates the total sales of the current quarter and the forecasted sales of the next quarter. The third box is dedicated to present the forecasting accuracy considering the past forecasting results obtained by the model, this box can give a better understanding of the reliability of the model's predictions.
The total sales trend section is located at the center of the dashboard and is a time series that shows the last 12 months sales compared with previous year total sales. This section has the goal to give to the managers a comprehensive overview about the sales results compared to the previous year.
The lower rectangle includes the revenues by sector, and aims to identify for the current period and for the next forecasted one how much, in terms of percentage, each sector contributes to the total sales of the ecommerce. It's possible to switch between the current month and the next month predictions through a button located on the top right of the box. There are three donut charts, one for each time frame of prediction: weekly, monthly, quarterly.
On the right, there's the top three sectors monthly forecast section, which aims to show the top three performing sectors in terms of sales considering the last three months and also considering the next two forecasted ones. For each sector there's its sales time series, and the predicted months are color coded with red.

In conclusion, forecasting the future sales of an ecommerce store can provide strategic advantages to the company that can be better managed through data driven decisions. In this project different forecasting models were deployed to predict the future sales of the ecommerce, and the best performing (in terms of MAPE) for each time frame was then used to deploy the sales forecast.

References

[1] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[2] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[3] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[4] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[5] David E. Rumelhart and James L. McClelland. "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. 1987, pp. 318–362.

[6] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.