Comparative Analysis of Sales Forecasting Models: A Multifaceted Approach and Empirical Study in a Grocery Retail company

Andrea Boscolo
Master degree in
Digital Transformation Management
University of Bologna
Email: andrea.boscolo@studio.unibo.it

***Abstract*—In the contemporary business landscape, accurate sales forecasting is pivotal for organizations navigating dynamic markets. This paper delves into the realm of sales data forecasting, specifically focusing on a comparative analysis of diverse models applied to real-world sales data from a Grocery Retailer. The study spans traditional statistical approaches, machine learning methodologies, and advanced deep learning techniques. The investigation reveals nuanced relationships between model complexity and predictive performance, challenging the assumption that more intricate models consistently outperform simpler ones. Surprisingly, a foundational baseline model, Holt-Winters Exponential Smoothing, surpasses more complex counterparts, emphasizing the importance of thoughtful model selection. However, the Multilayer Perceptron (MLP) model, incorporating exogenous variables, demonstrates superior performance, highlighting the significance of discerning challenging non-linear relationships.**

1. Introduction

In today's dynamic and competitive business landscape, accurate sales forecasting is crucial for organizations seeking to optimize their operations, allocate resources efficiently, and stay ahead of market trends. As businesses grapple with an ever-increasing volume of data, the ability to harness predictive analytics and sophisticated models becomes paramount. This paper delves into the realm of sales data forecasting, focusing on the comparative analysis of various forecasting models to discern the most effective approaches for predicting future sales trends.

Sales forecasting serves as the bedrock for strategic decision-making, aiding businesses in anticipating market demands, aligning inventory levels, and formulating targeted marketing strategies. With the advent of advanced analytics and machine learning techniques, the landscape of sales forecasting has undergone a transformative shift. However, the proliferation of diverse forecasting models has created a need for a comprehensive evaluation to discern the strengths, weaknesses, and contextual suitability of each.

This paper constitutes an investigation into the realm of sales data forecasting, specifically concentrating on the application of diverse regression models spanning traditional statistical approaches, machine learning methodologies, and advanced deep learning techniques. The empirical focus of this study is rooted in real-world sales data obtained from a Grocery Retailer (GDO).

2. Proposed Method

The methodology employed in this study unfolds across two pivotal stages: an initial data preprocessing and exploration, followed by the application of a diverse set of forecasting models.

The analysis delves into the application of various forecasting models, each selected for its distinct attributes. A baseline model was used to provide a fundamental benchmark for evaluation. The other used models are SARIMAX (and univariate ARIMA to select parameters), XGBoost and MLP. While the SARIMAX model encapsulates the temporal dynamics, the XGBoost algorithm brings forth the power of gradient boosting, and the Multilayer Perceptron (MLP) represents the deep learning paradigm. This multifaceted approach not only caters to the complexity of sales data but also enables a comprehensive comparison across different modeling paradigms. Through this combined methodology, I aim to extract insights into the forecasting landscape, offering a foundation for businesses seeking strategies to predict sales within the Grocery Retail domain.

2.1. Data preprocessing and exploration

The dataset has been extracted from a Grocery Retail's Data warehouse and is composed of monthly records from 01/01/2005 to 01/10/2022. Data contained in the record are aggregated for the total number of shops. Each record is composed of the sales amount (Euros), the sales quantity, the number of open shops in that month and an attribute that describes to which distribution channel that record belongs, i.e., "1" for Retail channel and "2" for Wholesale channel.

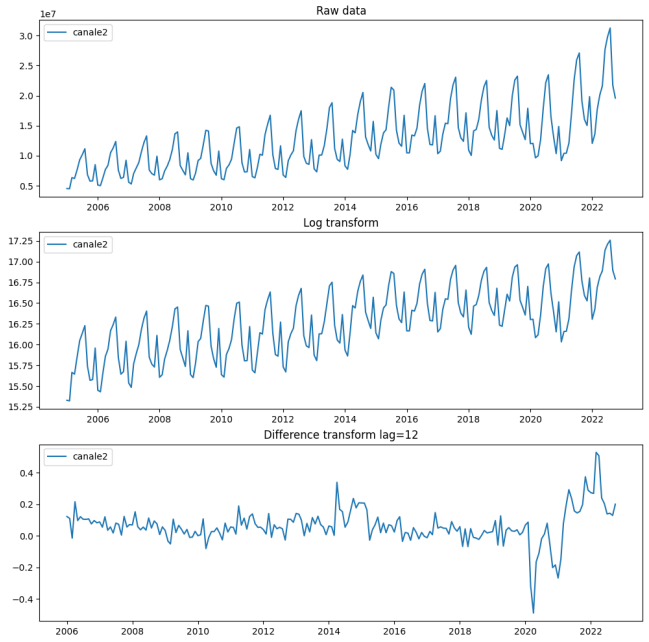

Figure 1. Raw, Logged and Differentiated data

I decided to concentrate the work on the Wholesale channel given that it has a peculiar trend and seasonality. The variable used in this work is the sales amount, given that quantities has been recorded with different methods over the years by the company.

I performed the following operations on the data:

- Log Transformation: The natural logarithm is applied to the original time series to mitigate the effect of trend, particularly useful when dealing with data exhibiting non-linear growth.
- Difference Transformation: The difference between each data point and the one at a fixed lag (in this case, 12 time units) is computed. This "difference transform" is commonly used to stabilize variance and achieve stationarity in the time series.

As it is notable from the Figure 1, stationarity is achieved differentiating with 12 time units the logged time series.

STL (Seasonal-Trend decomposition using LOESS) has then been used to decompose the time series into three main components: Seasonal, Trend, and Residual.

- Seasonal component represents the repeating patterns or seasonality in the time series, i.e., the recurring patterns that occur at regular intervals (e.g., daily, monthly, yearly).
- Trend component captures the long-term, slowly changing patterns in the time series, i.e., the overall direction or tendency of the data over an extended period.
- Residual component accounts for the random fluctuations or noise in the time series. After removing the seasonal and trend components, what remains is the residual, which ideally should exhibit no discernible pattern.

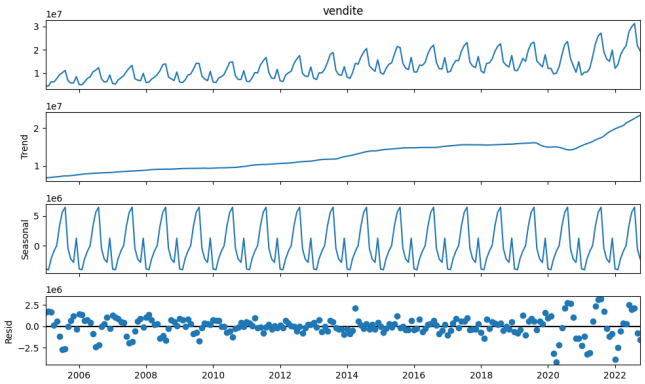

Figure 2. STL decomposition

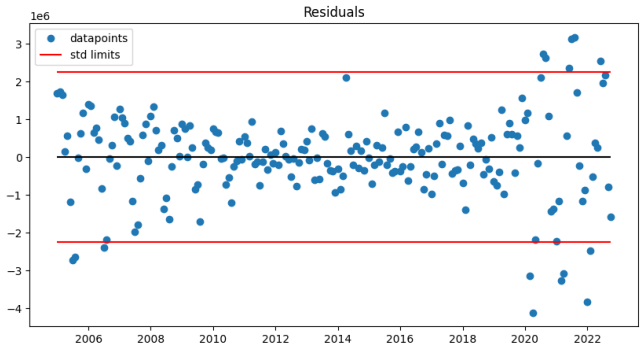

Figure 3. Residuals against a threshold

As shown in Figure 2, after decomposing the time series and removing trend and seasonal components, residuals are quite variable. In Figure 3 I plotted the residuals against a threshold, in which the upper boundary is calculated as the Standard Deviation of residuals multiplicated by 2 and the lower as the opposite. This plot shows that residuals contain some noise, particularly starting from 2020's records. This could be attributed to the unusual variation in sales that COVID-19 brought and the subsequent increase in inflation rates.

Then I tried to check if the attribute Number of shops could be useful as input for providing a better forecasting result by the models. A regression of deseasoned data on the number of shops has been performed to search for a possible correlation. I decided to use deseasoned data to exclude the strong seasonal component and to check if the trend component was mainly given by the increasing number of shops. The resulting r2 coefficient is 0.839, so particularly strong, as can be seen by Figure 4.

Given the strong correlation, I will use the number of shops as inputs in the following multivariate models.

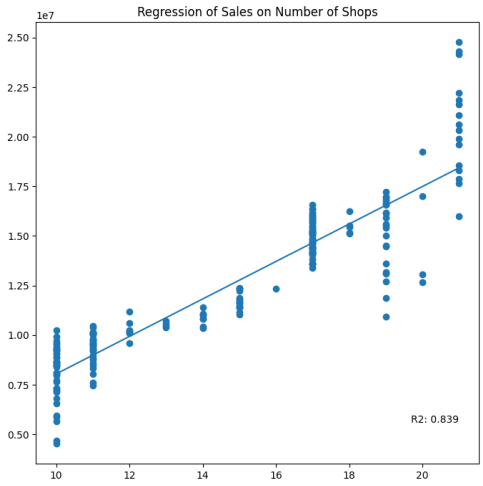

Figure 4. Residuals against a threshold

2.2. Forecasting models

The general training and forecasting approach was to split training and test sets to forecast the last 12 monthly sales values. In the sections below, I briefly explain each used model, providing some explanation regarding the chosen parameters and the construction of the input dataset, given that some of the models do not directly extrapolate seasonal and temporal information.

2.2.1. Foundational baseline model: Holt-Winters Exponential Smoothing. Holt-Winters Exponential Smoothing is the method employed as the foundational baseline model for evaluating the performance of the other more advanced forecasting techniques. Holt-Winters Exponential Smoothing is a time series forecasting approach that captures and extrapolates both trend and seasonality in the data. This method extends simple exponential smoothing by incorporating components for the level, trend, and seasonality of the time series, making it particularly suitable for datasets with systematic patterns over time. By leveraging the smoothing parameters for each component, the Holt-Winters method provides a flexible framework for capturing and predicting temporal variations.

2.2.2. SARIMAX. The first sophisticated forecasting model that was used is the Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX), which is an extension of the traditional ARIMA framework. Particularly, ARIMA (Autoregressive Integrated Moving Average) is a time series forecasting model that combines autoregressive (AR) and moving average (MA) components to capture temporal patterns. The model is denoted as ARIMA(p, d, q), where 'p' represents the order of the autoregressive component, 'd' is the order of differencing, and 'q' is the order of the moving average component. While ARIMA is effective for capturing basic temporal patterns, SARIMAX (Seasonal ARIMA with Exogenous Variables) enhances the model by introducing the capability to incorporate external factors, known as exogenous variables. SARIMAX is denoted as SARIMAX(p, d, q)(P, D, Q, s), where (P, D, Q) represent the seasonal orders, and 's' is the length of the seasonal cycle. I chose to strategically select hyperparameters for the SARIMAX by leveraging the automated selection capabilities of the autoarima function.

Specifically, my code executes an autoarima function only on sales values, without taking into consideration the number of shops, to extract the first and second (seasonal) order parameters performing a grid search. The parameters (p, d, q)(P, D, Q, s) that were automatically selected to input in the SARIMAX model were (1, 1, 1)(0, 1, 1, 12). Sales data were then used as the target variable, while considering the number of shops as an exogenous variable, recognizing its potential influence on sales patterns.

2.2.3. XGBoost Regressor. The second adopted model is XGBoost Regressor, which is a specific implementation of the XGBoost algorithm designed for regression tasks. XGBoost, short for eXtreme Gradient Boosting, is a powerful and efficient machine learning algorithm that belongs to the ensemble learning category. Leveraging a sequential construction of decision trees, XGBoost corrects errors iteratively, fostering a highly accurate predictive model. XGBoost Regressor is used when the target variable is continuous, and the algorithm is trained to predict numerical values.

The data has been prepared to be input into the model as a rolling window dataset, adding the actual number of shops. The rolling window dataset construction involves a methodical arrangement of historical sales data within a temporal context, facilitating sequential modeling and prediction. Utilizing a lookback period of 12 months, the dataset is structured to include lagged features, denoted as 't-1', 't-2', ..., 't-12', where each term corresponds to the sales data at a specific time point in the past. This approach captures the temporal dependencies inherent in the time series, enabling the model to consider the influence of historical observations on the current sales figures. The inclusion of the number of shops as an additional feature further enriches the dataset, accounting for potential external factors influencing sales dynamics. By systematically organizing the data in this rolling window format, the model is poised to learn and adapt to evolving patterns, contributing to the robustness of the forecasting methodology employed.

The chosen loss function to be minimized during the training of the XGBoost model was the mean squared error, which is a common metric for regression problems; while the number of estimators, i.e., the parameter that determines the number of trees (estimators) to be created in the ensemble during the training process, was set to 1000, meaning that the XGBoost model will consist of 1000 decision trees.

2.2.4. MLP. The third and last model is the Multilayer Perceptron (MLP). MLP represents a class of artificial neural networks characterized by multiple layers of interconnected nodes, each layer contributing to the extraction of complex

	MAE
Baseline	2,128,376
autoARIMA	2,555,850
SARIMAX	3,416,919
XGBoost	3,180,776
MLP	1,870,494

Figure 5. MAEs

features from input data. MLP is a versatile architecture commonly employed for regression tasks due to its capacity to model intricate non-linear relationships within datasets. Comprising an input layer, one or more hidden layers, and an output layer, MLP leverages a systematic process of forward and backward propagation to learn and optimize weights, adjusting the network's parameters during training to minimize prediction errors.

The neural network model utilized in this study is structured as a sequential architecture using the Keras library. Comprising a single hidden layer, the model incorporates rectified linear unit (ReLU) activation functions. The number of neurons in this hidden layer is determined by the heuristic formula $n = TestLenght * 2 + 1$, a common approach for establishing an appropriate number of neurons in a single-layer perceptron. The output layer consists of a single neuron, aligning with the regression nature of the task. The model is compiled with the mean absolute error as the loss function and the Adam optimizer. This sequential configuration aims to capture intricate patterns and relationships within the dataset, making it well-suited for regression tasks with complex and non-linear dependencies.

Here, the construction of the input dataset was constructed, as for the XGBoost model, as a rolling window dataset but instead of rolling only the sales even the variable relative to the number of shops was lagged. This means that the dimension of the input dataset was 24, consisting of 12 lagged sales values, and 12 lagged numbers of shops.

3. Results

In assessing the predictive performance of the different used models, the Mean Absolute Error (MAE) serves as a key metric for accuracy in regression tasks. Surprisingly, the baseline model, i.e., Holt-Winters Exponential Smoothing, outperforms more complex counterparts, achieving a notably lower MAE of 2,128,376. The AutoARIMA model, despite its automated algorithmic sophistication, yields a slightly higher MAE at 2,555,850. Furthermore, models incorporating additional complexities, such as seasonality and exogenous variables like SARIMAX (MAE: 3,416,919) and XGBoost (MAE: 3,180,776), demonstrate increased predictive challenges in capturing the intricate sales dynamics. Intriguingly, the Multilayer Perceptron (MLP) model, a neural network architecture renowned for its capacity to handle non-linear relationships, outshines all other models with the lowest MAE at 1,870,494. The surprising efficacy

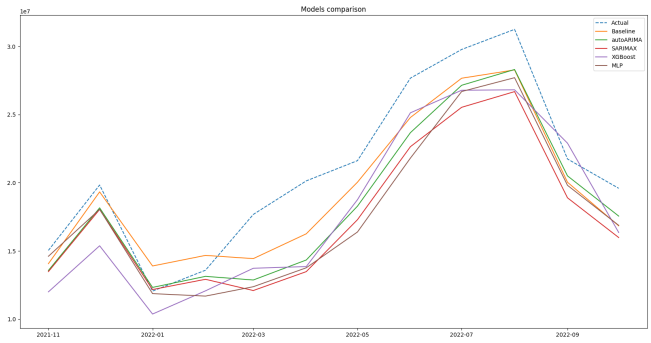

Figure 6. Models' forecasted sales

of the baseline model may stem from its ability to capture fundamental trends and patterns inherent in the sales data without succumbing to unnecessary complexity—a noteworthy observation that underscores the importance of thoughtful model selection and the potential for simpler models to outperform in certain contexts.

Comparing models against the mean sales value (20,833,356) reveals that the baseline and AutoARIMA models closely align with the average, demonstrating effective predictions. Models with increased complexity, like SARIMAX and XGBoost, show more deviation. Notably, the Multilayer Perceptron (MLP) outperforms others, minimizing errors relative to the mean. This emphasizes the nuanced trade-off between model complexity and predictive accuracy, showcasing the effectiveness of simpler models in capturing essential sales trends or the need to exploit exogenous variables with more complex models in order to capture the difficult non-linear combination that intercourse with the target variable.

As depicted in Figure 6, it becomes evident that all models tend to underestimate the peak observed in the latter portion of the test set. A potential explanation for this discrepancy is discernible when examining Figure 1 and 2, where the presence of outliers in the final segment of the time series is apparent. This period is characterized by lower minimums and higher maximums, which may contribute to the models' difficulty in accurately capturing the peak values during this particular phase.

4. Conclusions

The analysis of various models for forecasting sales in a Grocery Retail company reveals a nuanced relationship between model complexity and predictive performance. Contrary to the expectation that more intricate models consistently yield superior results, this study highlights instances where a simpler baseline model outperforms its more complex counterparts. The baseline model, relying solely on the target variable, demonstrates that sophistication isn't always synonymous with accuracy. However, the MLP model's performance underscores the importance of incorporating exogenous variables. It suggests that while such variables can enhance predictive capability, the model must

adeptly discern challenging non-linear relationships inherent between independent and dependent variables for optimal performance.

Further enhancements can be pursued to augment the predictive capabilities of the presented models. Firstly, a more refined exploration of optimal hyperparameters could be undertaken, incorporating a validation step dedicated to identifying parameters that best suit our specific problem. Secondly, the integration of additional exogenous variables could provide deeper insights into new relationships within the data, such as incorporating monthly inflation rates. Thirdly, an intriguing avenue involves the combination of different models in a unified pipeline, leveraging the strengths of each. For instance, employing a univariate baseline or SARIMA model to capture temporal dependencies in the target variable, while employing a more intricate model like MLP to discern deterministic patterns within the residuals, potentially linked to exogenous variables.

References

[1] Shumway, Robert H., et al. "ARIMA models." Time series analysis and its applications: with R examples (2017): 75-163.

[2] Vagropoulos, Stylianos I., et al. "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting." 2016 IEEE international energy conference (ENERGYCON). IEEE, 2016.

[3] Kalekar, Prajakta S. "Time series forecasting using holt-winters exponential smoothing." Kanwal Rekhi school of information Technology 4329008.13 (2004): 1-13.

[4] Avanijaa, J. "Prediction of house price using xgboost regression algorithm." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.2 (2021): 2151-2155.

[5] Shiblee, Md, Prem Kumar Kalra, and B. Chandra. "Time series prediction with multilayer perceptron (MLP): A new generalized error based approach." Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25-28, 2008, Revised Selected Papers, Part II 15. Springer Berlin Heidelberg, 2009.