The wisdom of crowds in forecasting at high-frequency for multiple time horizons: A case study of the Brazilian retail sales

Gustavo Picoli Lopes[†]

Abstract This case study compares the forecasting accuracy obtained for four daily Brazilian retail sales indices across four prediction horizons. The performance of traditional time series forecasting models, artificial neural network architectures and machine learning algorithms were compared in order to evaluate the existence of a single best-performing model. Subsequently, ensemble methods were included in the model comparison to determine if accuracy could be improved. Evidence from this case study suggests that a consistent forecasting strategy exists for the Brazilian retail indices by applying both seasonality treatment for holidays and calendar effects and by using an ensemble method whose main inputs are the predictions of all models with calendar variables. This strategy was consistent across all 16 index and time horizon combinations, as ensemble methods either outperformed the best single models or showed no statistical difference from them in a Diebold-Mariano test.

Keywords: Forecasting; Retail; Artificial neural networks; Machine learning; Ensemble methods.
JEL Code: C22, C45, C52, C53.

1. Introduction

According to the Brazilian Society of Retail and Consumption (SBVC), "retail" refers to any economic activity involving the sale of a good or service to the final consumer. In the composition of the Gross Domestic Product (GDP), retail is related to household consumption, which accounted for 62.9% of the Brazilian GDP in 2020, as per data from the system of national accounts released by the Brazilian Institute of Geography and Statistics (IBGE).

The payments market is among the fastest-growing sectors in Brazilian retail. According to the balance sheet released by the Brazilian Association of Credit Card and Services Companies (ABECS), the total value of card transactions reached R$ 2.65 trillion in 2021, marking a 33.1% growth from

Submitted on November 30, 2021. Revised on March 12, 2022. Accepted on April 4, 2022. Published online in June 2022. Editor in charge: Marcelo Fernandes.
[†]Cielo, Brazil: picoli.gustavo@gmail.com

©2022 Sociedade Brasileira de Finanças, under a Creative Commons Attribution 3.0 license

the previous year and constituting 54.6% of household consumption in the third quarter. The indices analyzed in this study are composed by the total monetary value of card transactions captured by a single Brazilian acquiring company, categorized into **total retail**, **durable goods**, **non-durable goods** and **services**.

The nature of the data permits daily evaluation of these indices. Thus, events that are representative of retail calendar, such as holidays, can be studied independently from regular sales days. This seasonality dynamic is known as calendar effect and generates an abrupt increase in the volatility of daily sales indices.

This study assesses the accuracy performance of both conventional time series prediction models—Naïve, Holt-Winters, TBATS and SARIMA—and popular machine leaning and artificial neural network algorithms—Multilayer Perceptron (MLP), Long-Short Term Memory (LSTM), eXtreme Gradient Boosting (XGBoost) and Facebook's Prophet model—in a forecasting scenario across multiple time horizons. All these models, with Naïve being the only exception, depend on hyperparameters, which directly affect predictions.

For hyperparameter selection, a simple holdout approach was employed, wherein the in-sample period was split into two smaller sections: one for hyperparameter estimation and another for validation. While authors like Arlot and Celisse (2010), Bergmeir and Benítez (2012), Bergmeir et al. (2014) and Pinto and Marçal (2021) have reported promising accuracy results when forecasting time series using a cross-validation approach, the consistency of these results remains a debated topic in academic literature. A cross-validation method proposed by Burman et al. (1994) indicated that the h-block approach was asymptotically inconsistent, contrasting with findings by Shao (1993), who recommended this model selection approach over simple leave-one-out cross-validation. Therefore, Racine (2000) introduced a modified "hv-block" cross-validation, suggesting that the method was asymptotically consistent. However, a recent note by Zheng (2019) highlighted an error in Racine's work, leaving the consistency of "hv-block" cross-validation open to discussion. Notably, in Brazilian retail sales forecasting literature, authors like Pasquotto (2011), Angelo et al. (2011) and Bessa (2018) have utilized either the Akaike or Schwarz information criteria for hyperparameter selection in traditional time series models such as SARIMA.

Forecasts were produced for four prediction horizons: the following day ($D+1$), after seven consecutive days ($D+7$), after thirty consecutive days ($D+30$), and after ninety consecutive days ($D+90$). The Root Mean Squared Error (RMSE) was iteratively computed in the out-of-sample period for each

scenario, with models being re-estimated after each prediction cycle. This methodology was devised to emulate real-world situations where an economic agent updates predictions daily.

An open discussion in time series forecasting literature is related to the conditions under which neural networks and other machine learning algorithms outperform traditional models. In the 1990s, with the advancement of neural network research efforts, Gorr (1994) and Hill et al. (1994) advocated for future studies examining scenarios where an artificial neural network might surpass traditional model. In Brazil, authors like Pasquotto (2011), Angelo et al. (2011) and Bessa (2018) conducted forecasting studies using neural network models for retail sectors; however not all attained significant accuracy improvements. The recent doctoral thesis by Bessa (2018) incorporated the XGBoost algorithm, which outshone other models for the Brazilian fashion retail industry. Her work also reported accuracy gains by applying ensemble methods. This paper evaluates her recommendation but in a different context—by studying the Brazilian retail sector as a whole and forecasting over various time horizons.

Therefore, the aims of this case study are: (1) assessing whether a single model can predict daily indices with greater accuracy, juxtaposing traditional models with artificial neural networks and machine learning algorithms; (2) determining the consistency of results across the four sales indices; (3) identifying whether the results are consistent across prediction horizons; (4) gauging the impact of modeling the calendar effect on accuracy; (4) evaluating if an ensemble approach enhances forecasting outcomes; and (5) exploring the feasibility of devising a reliable forecasting strategy for Brazilian retail sales data.

An insightful way to encapsulate the primary conclusion of this study is to draw parallels with James Surowiecki's renowned theory "Wisdom of Crowds". According to his theory, collective opinions of groups often surpass those of individual experts. In line with this, for all indices and prediction time horizons, an ensemble blending method—employing all available models with calendar variables—either outperforms the single best model, or matches it statistically. Therefore, a dominant strategy to enhance accuracy for forecasting high-frequency Brazilian retail data over multiple time horizons has been identified: (1) adjust the time series for seasonality by examining holidays and calendar effects; (2) generate predictions for the indices and their log transformations using traditional models, artificial neural networks, and machine learning algorithms; and, finally, (3) ensemble all these predictions by taking either the mean or median to produce a singular forecast.

The study unfolds over five sections. The subsequent section offers a concise review of both retail sales forecasting literature and the history of time series forecasting. Section 3 delves into the methodology, while Section 4 presents the forecasting results. Finally, Section 5 summarizes the conclusion for each objective of this article.

2. Related literature and motivation

The topic of retail sales forecasting has been extensively studied in the academic literature, both in business administration——for supply chain optimization and strategic planning purposes——and in macroeconomics, which examines the aggregate sales for a country's economic growth dynamics.

Authors such as Geurts and Kelly (1986), Frank et al. (2003), and Au et al. (2008) have studied specific sectors within retail, detailing the outcomes achieved when comparing various forecasting models. For example, Geurts and Kelly (1986) found that an exponential smoothing model could produce superior forecasting results compared to a SARIMA model. However, Frank et al. (2003) and Au et al. (2008) reported better outcomes for the fashion industry using artificial neural networks (ANNs).

By analyzing the forecasting problem for the retail sales aggregate of an economy, Alon et al. (2001) discovered interesting results using artificial neural networks when compared to traditional ARIMA and exponential smoothing models. The authors attributed these results to the neural networks' capability to capture non-linear patterns in the time series, thereby providing enhanced accuracy. Afterwards, authors like Chu and Zhang (2003) and Aye et al. (2015) also assessed the performance of ANNs against traditional models. Chu and Zhang (2003) achieved superior MAPE and RMSE results using ANNs on deseasonalized data, while the best results in Aye et al. (2015) came from merging different models for forecasting. Notably, Aye et al. (2015) also explored forecasting across various prediction horizons, obtaining similar results throughout.

Using Brazilian retail sales data, Almeida and Passari (2006) compared forecasting models for monthly SKU's sales data from a prominent retail corporation. They reported superior forecasting results in RMSE and MAPE using ANNs, attributing these findings to the ANNs' ability to discern interactions between products, simulating demand's cross elasticity. Bessa (2018) also obtained interesting results employing the LSTM neural network model on weekly sales data from a specific retailer in the fashion industry. Her results showed up to a 54.32% increase in accuracy when contrasting the LSTM's performance with the method previously employed by the retail

company.

However, not all researchers in Brazil have achieved consistent results. For example, Pasquotto (2011), who analyzed monthly aggregated sales from the pharmaceutical, fertilizer, and air traffic sectors using Elman Networks (a particular neural network architecture), found that this model couldn't surpass a straightforward SARIMA model. Similarly, Angelo et al. (2011), who studied forecasting models using aggregated retail sales from Brazil's Monthly Survey of Trade (PMC), didn't observe any significant forecasting gains with neural networks.

Other significant works in the forecasting and model hyperparameter selection literature include studies by Swanson and White (1997), Hippert et al. (2005), Hyndman et al. (2002), and De Gooijer and Hyndman (2006). These papers encompass both traditional models, like exponential smoothing and SARIMA, as well as neural network structures. Researchers such as Zadeh et al. (2010), Glorot and Bengio (2010), Nair and Hinton (2010), Karlik and Olgac (2011), Glorot et al. (2011), and Maas et al. (2013) delved into different activation functions for neural network performance, while Ruder (2016) evaluated various learning rate optimizers.

Since its introduction by Chen and Guestrin (2016) as a machine learning system rooted in the gradient-boosted tree algorithm, the eXtreme Gradient Boosting (XGBoost) has been reported to yield superior forecasting outcomes, as evidenced by studies from Saradhi and Naresh (2018), Ji et al. (2019), and Bessa (2021). Saradhi and Naresh (2018) also tested this algorithm on daily frequency time series, matching the frequency of the indices assessed in this case study.

Another algorithm incorporated into this study's model comparison is a toolset developed by Facebook researchers, termed "Prophet" Taylor and Letham (2018). The creators of this algorithm assert that Prophet is robust to noise and outliers in data, and excels in time series that exhibit pronounced seasonal effects and multiple seasonality patterns, aligning with the indices discussed in this study. Since its release, researchers like Fang et al. (2019) have reported enhanced accuracy using Prophet when evaluating the Morgan Taiwan index in the financial market. Toharudin et al. (2021) compared this model against the LSTM neural network for air temperature forecasting, with Prophet surpassing the neural network under specific conditions. Lastly, Garlapati et al. (2021) compared both Prophet and ARIMA models for stock price prediction.

Regarding the appropriate metric for prediction error, authors like Chatfield (1988), Armstrong and Collopy (1992) and Makridakis et al. (1998)

highlighted the pitfalls of choosing inappropriate metrics for evaluating model performance. They concluded that RMSE is unsuitable for comparing different time series (since they are scale-dependent), and that MAPE has limitations when assessing values close to zero. Therefore, the RMSE metric was selected for the Brazilian retail sales indices, as comparing various time series is not a primary focus of this study.

The following points outline the contributions of this article to the literature: (1) It assesses a variety of time series forecasting models, from traditional approaches to intricate machine learning algorithms; (2) The indices featured in this paper represent sales from the entirety of Brazilian retail, not just a specific sector, facilitating comparison with Pasquotto (2011); (3) The study examines high-frequency data, which tends to be noisy, as demonstrated in Frank et al. (2003); (4) Various prediction horizons are evaluated, using the research of Aye et al. (2015) as a benchmark; and, finally, (5) The paper tests the hypothesis, put forth by Bessa (2021), that her findings for ensemble methods and seasonality treatments can be extended to sectors beyond fashion retail.

3. Dataset and methodology

We begin by introducing the dataset used in this case study and describing its construction. The experimental methodology is divided into three sections: Section 3.2 outlines the evaluated models, Section 3.3 details the tested hyperparameters, and Section 3.4 explains both the holdout method and out-of-sample forecasting procedures.

3.1 Dataset

The time series for this study come from the aggregation of all card transactions recorded by a Brazilian acquiring company between January 1st, 2014 and July 31st, 2018—a total of 1673 daily observations. This data represents the monetary sum of all transactions between cardholders and commercial establishments in Brazilian retail in which this acquiring company operated. The data was transformed to base 100 index—with January 2nd, 2014 as the index reference—and was divided in four distinct time series, as follows:

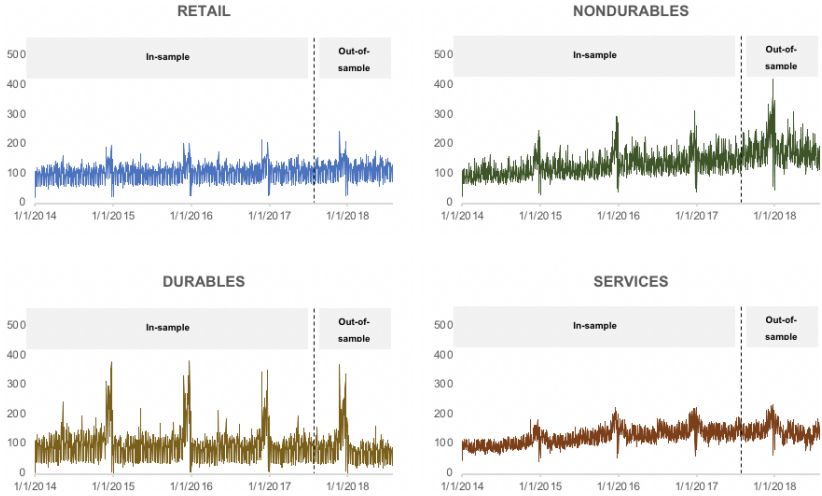
(a) Total retail (RETAIL): aggregated nominal revenue index for all establishments;
(b) Non-durable goods (NONDURABLES): aggregated nominal revenue index for establishments in the sectors of non-durable and semi-durable goods;

(c) Durable goods (DURABLES): aggregated nominal revenue index for establishments in the sector of durable goods;
(d) Services (SERVICES): service-providing establishments' aggregated nominal revenue index.

Each index was split into in-sample and out-of-sample periods. The in-sample period spans from January 1st, 2014 to July 31st, 2017, totaling 1308 observations. The out-of-sample period ranges from August 1st, 2017 to July 31st, 2018, in a total of 365 observations. We divided the time series in this manner to ensure that the out-of-sample period contains all days from a complete year, enabling a comprehensive analysis of the calendar effect.

The retail indices and their division between in-sample and out-of-sample periods are shown in Figure 1.

Figure 1
Time series for each retail index divided in in-sample and out-of-sample periods



All four indices exhibit a strong seasonal pattern. The most significant partial correlations are found up to the seventh lag and, subsequently, at lags multiple of seven, indicating strong seasonality within the days of the week. The summary statistics, along with the autocorrelation function (ACF) and partial autocorrelation function (PACF), are provided in the Appendix.

Given that NONDURABLES, DURABLES and SERVICES indices aggregate data from a diverse range of retail sectors, from drugstores to airlines, it is worth describing them in detail. Table 1 lists all these Brazilian retail

sectors which are present in the data.

Table 1
Retail sectors that compose the NONDURABLES, DURABLES and SERVICES indices

NONDURABLES	DURABLES	SERVICES
Cosmetics and personal hygiene	Sportive articles	Travel agencies
Drugstores and pharmacies	Electronics	Bars and restaurants
Bookstores and stationery stores	Department store	Vehicle rent
Gas stations	Construction materials	Automotive services
Supermarkets and hypermarkets	Furniture and decoration	Airline companies
Specialized food retail	Jewelry	Hairdressers
Veterinarians and pet shops	Clothing	Hotels
		Entertainment and leisure
		Transportation

Another important feature shared by the four time series is the presence of pronounced peaks and troughs on certain days of the year. Table 2 pinpoints the exact dates of the highest and lowest sales for each index:

Table 2
Highest and lowest retail sales days by index

	lowest		highest	
Index	value	day	value	day
RETAIL	21.16	01-01-2014	245.14	11-24-2017
NONDURABLES	23.61	01-01-2014	424.20	12-23-2017
DURABLES	4.33	01-01-2014	384.39	12-23-2015
SERVICES	40.59	12-25-2014	236.18	12-22-2017

The seasonal patterns clearly show that Christmas Day and the first day of the year have the lowest retail revenue, whereas the shopping days before Christmas and Black Friday are the ones with highest revenue.

As a primary goal of this case study is to scrutinize the impact of modeling the retail calendar effect, we incorporated some control variables into the dataset to isolate its influence on sales. Several evaluated models accommodated the inclusion of exogenous variables. Thus, we tested models both with and without these control variables, which are presented in Table 3.

Table 3
Control variables used for retail's calendar effect

variable	description
C_WORLD_DAY	World Day of Peace
C_CARNIVAL	Brazilian's carnival period
C_TIRADENTES	Tiradentes' Day
C_EASTER	Easter Day
C_CORPUS_CHRISTI	Feast of Corpus Christi
C_WORK	'Work day' holiday
C_INDEPENDENCE	Brazilian's Independence Day
C_DAY_DEAD	Day of the Dead
C_REPUBLIC	Proclamation of the Republic
C_MOTHER	Mother's Day
C_VALENTINE	Valentine's Day
C_FATHER	Father's Day
C_CHILDREN	Children's Day
C_CHRISTMAS	Christmas Day
C_BLACK_FRIDAY	Black Friday
C_A_MOTHER	seven days before Mother's Day
C_A_VALENTINE	seven days before Valentine's Day
C_A_FATHER	seven days before Father's Day
C_A_CHILDREN	seven days before Children's Day
C_A_CHRISTMAS	seven days before Christmas

It is important to notice that seasonality pattern can be partially captured by some models without the necessity of adding a control variable (e.g. C_A_FATHER and C_A_CHILDREN, both occurs in the same period of the year). However, the dataset has a limited number of complete yearly cycles; for instance, the months from August to December appear only four times, with one occurrence in the out-of-sample period. The feature engineering, facilitated by these control variables, aims to enhance accuracy given this data constraint.

Finally, it's important to emphasize that the seasonality modeling approach employed in this case study diverges from that of Bessa (2021). In her work, differentiation and polynomial methods were used to extract seasonality patterns from the time series. While these techniques could be applied to our dataset, they might pose limitations, especially given that some holidays, such as Carnival and Corpus Christi, do not recur consistently within the same time frame each year.

3.2 Experimental Setup 1: Model descriptions

Twenty-one different models were tested, ranging from traditional time series methods, artificial neural network architectures, machine learning algorithms, to ensemble methods. Table 4 provides a brief description of each.

Table 4
Model types used in the forecasting procedure

#	model	description
1	NAIVE	Naïve forecasting method
2	SEASONAL_NAIVE	Seasonal Naïve forecasting method
3	HOLT_WINTERS	Triple exponential smoothing model
4	TBATS	Exponencial smoothing state space model
5	SARIMA	SARIMA model
6	SARIMA_LOG	SARIMA model and logarithm treatment
7	MLP	Feedforward neural network Multilayer Perceptron
8	LSTM	Recurrent neural network Long Short-Term Memory
9	PROPHET	Facebook's time series forecasting algorithm
10	XGBOOST	eXtreme Gradient Boosting algorithm
11	ETS_CALENDAR	ETS model with control variables
12	ETS_LOG_CALENDAR	ETS model with control variables and logarithm treatment
13	SARIMA_CALENDAR	SARIMA model with control variables
14	SARIMA_LOG_CALENDAR	SARIMA model with control variables and logarithm treatment
15	MLP_CALENDAR	MLP model with control variables
16	LSTM_CALENDAR	LSTM model with control variables
17	PROPHET_CALENDAR	Prophet model with control variables
18	PROPHET_LOG_CALENDAR	Prophet model with control variables and logarithm treatment
19	XGBOOST_CALENDAR	XGBoost model with control variables
20	ENSEMBLE_BLENDING (MEAN)	Mean forecast from models 11 to 19
21	ENSEMBLE_BLENDING (MEDIAN)	Median forecast from models 11 to 19

It is possible to observe that three "types" of models were constructed for this study: (1) the "pure" ones, applied directly to each index; (2) the "logarithm" ones, where a logarithm transformation is applied to the time series and inverse transformed after the forecasting procedure; and (3) the "calendar" ones, which incorporate all the dummy variables described in Table 3.

The logarithm transformation was applied to models that presented at least one linear component in their structure (ETS, SARIMA and Prophet) so that some of the time series nonlinearity could be easily modeled. The main hypothesis is that this simple transformation could improve the accuracy of predictions. This transformation was feasible due to the nature of the indices, which are strictly positive (representing monetary aggregation of confirmed card transactions in Brazilian retail).

The Naïve methods are named as such because they consider only the last observation for forecasting. Equations (1) and (2) describe both the Naïve and

Seasonal Naïve methods, respectively:

$$\widehat{y}_{t+1} = y_t \tag{1}$$

$$\widehat{y}_{t+1} = y_{t-s} \tag{2}$$

where $\widehat{y}_{t+1}$ is the forecasted value for period $t+1$; y_t is the observed value in the most recent period; y_{t-s} is the observed value in the last seasonal period.

The seasonal period s considered in this study was a time lag of seven periods since the day-of-the-week pattern proved relevant in the PACF analysis. Although both methods appear simple at first, they serve as baseline models for accuracy results and can even generate satisfactory results in some conditions, according to Makridakis et al. (1982), Makridakis et al. (1993) and Lawrence et al. (2000).

Models like Holt-Winters, TBATS, SARIMA, Prophet, XGBoost, MLP and LSTM rely on specific hyperparameters and structures, which are covered in the next section. It is important to mention that the LSTM structure adopted in this study encompasses all its components, as initially proposed by Hochreiter and Schmidhuber (1997) and later refined by Gers et al. (2000).

3.3 Experimental Setup 2: Hyperparameters tested

As described in the last subsection, models are hyperparameter-dependent and it is necessary to establish a range of values for them. Table 5 shows all the values tested for each hyperparameter and the total number tested by model type. Some of them do not assume numerical value, but define a structure, as is the case of Holt-Winters and ETS methods (trend and seasonality components) and the activation functions and optimizers for both MLP and LSTM neural networks architectures.

In total, 3372 different combinations of hyperparameters were tested for each of the four retail sales indices. They were chosen as follows:

(a) Holt-Winters: The structure for trend and seasonality components follows the most common methods in literature, according to De Gooijer and Hyndman (2006);
(b) TBATS: The main advantage of this model is the inclusion of multiple and complex seasonality patterns presented in the time series. Therefore, since the indices are at daily frequency, the weekly, monthly and yearly seasonality were added to the seasonal cycle hyperparameter. A box-cox transformation was tested as described in De Livera et al. (2011);
(c) SARIMA, SARIMA_LOG, SARIMA_CALENDAR, SARIMA_LOG_CALENDAR: At least one of the integration parameters, d or D, assume the unit

Table 5
Hyperparameters tested by model type

model	hyperparameter	tested values	total
NAIVE			1
SEASONAL NAIVE			1
HOLT-WINTERS	seasonal cycle trend component seasonal component	7;14;21;28;35 additive additive; multiplicative	10
TBATS	seasonal cycle box-cox transformation damped trend	7;30.5;360.25 yes; no yes; no	4
SARIMA, SARIMA_LOG, SARIMA_CALENDAR and SARIMA_LOG_CALENDAR	p d q P D Q S	1;2;3;4;5;6 0;1 0;1 1;2;3;4;5 0;1 0;1 7	360 each
ETS_CALENDAR and ETS_LOG_CALENDAR	seasonal component ETS structure	7; (7,35); (7,35,364) ANA; AAA; AAdA; MNA; MAA; MAdA; MNM; MAM; MAdM; MMM; MMdM; ANM; AAM; AAdM; MMA; MMdA; AMA; AMdA; AMM; AMdM	60 each
PROPHET, PROPHET_CALENDAR and PROPHET_LOG_CALENDAR	changepoint scale seasonality scale seasonality mode	0.001;0.01;0.1;0.5 0.01;0.1;1;10 additive; multiplicative	32 each
MLP and MLP_CALENDAR	number of lags neurons in hidden layer number of hidden layers learning rate activation function optimizer number of epochs batch size	7;14;21;28;35 7;14;21;28;35 1 0.001;0.01;0.1 ReLU; sigmoid Adam; RMSprop 300 200	300 each
LSTM and LSTM_CALENDAR	number of lags neurons in hidden layer number of hidden layers learning rate optimizer number of epochs batch size	7;14;21;28;35 7;14;21;28;35 1 0.001;0.01;0.1 Adam; RMSprop 300 200	150 each
XGBOOST and XGBOOST_CALENDAR	learning rate max depth min child weight gamma subsample	0.02;0.04;0.06;0.08;0.1 2;4;6;8;10 1;10;50;100 0;10;100 0.5;1	600 each
		Total	**3372**

value to induce stationarity. The inclusion of moving average parameters, q and Q, was tested to check for gains in accuracy when modeling possible short-term shocks. The autoregressive parameter p, finally, was tested until the 6th lag, according to the results found in the PACFs;

(d) ETS_CALENDAR, ETS_LOG_CALENDAR: The "ETS structure" that was tested in these models followed the classification proposed by Hyndman et al. (2008), dividing ETS models in five classes. All model types from their classification were tested except for the non-seasonal ones, since all indices present strong seasonal patterns for reasons previously discussed in the text.

(e) PROPHET, PROPHET_CALENDAR, PROPHET_LOG_CALENDAR: Facebook's prophet model is composed by a set of different hyperparameters, which include, for example, the Fourier series frequency for seasonality. However, the algorithm's creators recommend tuning changepoint scale and seasonality scale, for both trend and seasonality. The impact these hyperparameters have on modelling can be found in Taylor and Letham (2018).

(f) MLP, MLP_CALENDAR: According to Swanson and White (1997), choosing a simpler neural network structure, with only one hidden layer, can bring good results, especially when sample size can be an issue. A combination of the possible values for the number of neurons in a hidden layer varied according to the number of lags tested for this model. Activation functions ReLU and sigmoid were selected for non-linearity induction. Zadeh et al. (2010), Glorot and Bengio (2010) and Karlik and Olgac (2011) have obtained interesting results with sigmoid function while Glorot et al. (2011) and Maas et al. (2013) found expressive convergence improvement with ReLU function. Finally, Adam and RMSprop optimizers were chosen—following the study made by Ruder (2016)—with three different learning rates (0.001 / 0.01 / 0.1);

(g) LSTM, LSTM_CALENDAR: The hyperparameter selection is similar to MLP and MLP_CALENDAR models with the difference that, in the LSTM model type, the network's activation function is already established within the memory block, as can be seeing in Gers et al. (2000);

(h) XGBOOST, XGBOOST_CALENDAR: The selected range of hyperparameters followed Chen and Guestrin (2016) recommendation for tuning, since boosting tree models can generate relatively complex models, which can easily overfit when used in out-of-sample data.

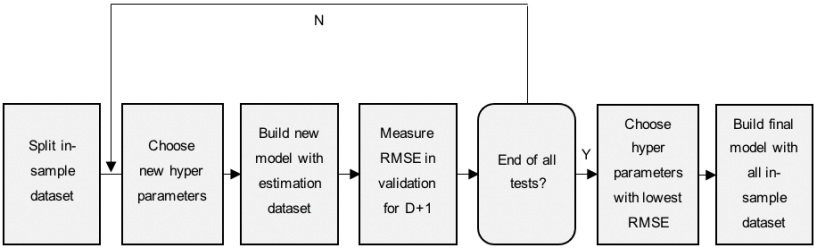
3.4 Experimental Setup 3: Hyperparameter selection and forecasting

In this study, we used a holdout method for hyperparameter selection before forecasting the indices for the out-of-sample period. This technique involves splitting the in-sample period into two subsets of data: one for model

estimation and the other for validation.

Therefore, the in-sample period was divided as follows: 1046 observations (the first 80% of the in-sample period) were used for model estimation, and 262 observations (the last 20% of the in-sample period) for validation. Each hyperparameter combination, defined in the previous section, had its performance tested. The final hyperparameters were chosen by selecting the combination with the lowest RMSE in validation dataset for $D+1$ forecasting horizon. Figure 2 summarizes the described methodology:

Figure 2
Flowchart for hyperparameter selection with the holdout procedure



One of the objectives of this case study, as described in the Introduction, is to compare the model's accuracy across different periods. To this end, we constructed four different forecasting horizons for the out-of-sample period, as described in Table 6:

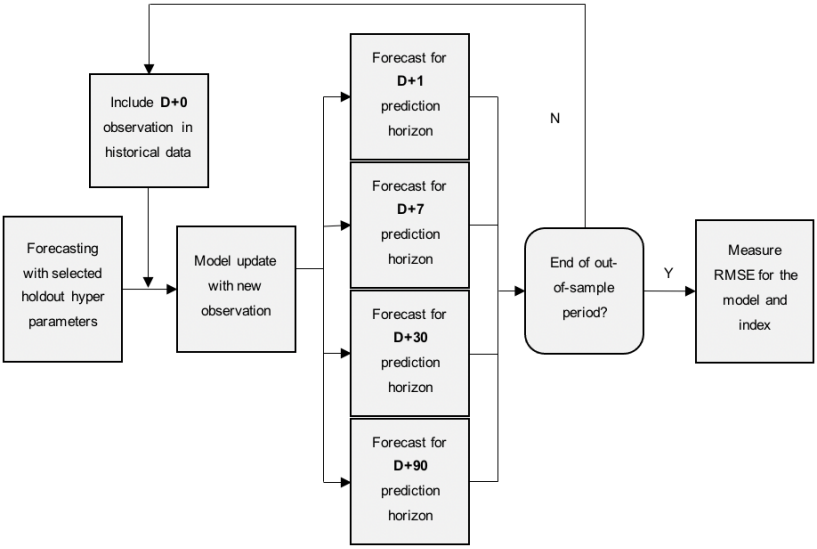
Table 6
Description for each forecasting horizon

horizon	description
D+1	forecasting the index's value for the next day
D+7	forecasting the index's value after seven days
D+30	forecasting the index's value after thirty days
D+90	forecasting the index's value after ninety days

After selecting the final hyperparameters for each model type and index using the holdout approach, we conducted the out-of-sample forecasting procedure. In this case study, to simulate a situation where an economic agent can update forecasts daily, a process of feedback loop was adopted by an expanding window approach. In other words, after forecasting one period ahead, the observation for that day was added to the historical data, and the models were

re-estimated before forecasting the subsequent period. Figure 3 provides an overview of the methodology:

Figure 3
Flowchart for the forecasting procedure



After forecasting all indices across all time prediction horizons, we constructed the ensemble blending models by taking the mean and median values of all available calendar models (models 11 to 19 from Table 4).

We conducted three distinct analyses to address the primary objectives of the case study: (1) Evaluation of non-calendar controlled models to determine if there exists a consistent model capable of forecasting with higher accuracy without seasonality treatment; (2) Evaluation of all models to verify whether there is a consistent error reduction by applying the seasonality treatment for calendar effect, and whether there is a model type that outperforms all others across all index and prediction horizons; and (3) Evaluation of ensemble models in comparison to individual models.

4. Forecasting results

This section is subdivided into **hold-out procedure results**, which exhibits the RMSE dispersion obtained during hyperparameter selection, and **out-of-sample forecasting**, which presents the forecasting results for each of

the three analyses described in the previous section.

4.1 Hold-out procedure results

As detailed in Table 5, a total of 3372 hyperparameter combinations were tested for each retail sales index using a hold-out approach. The combination with the lowest RMSE for $D+1$ forecasting period in the validation dataset was selected as the forecasting model for the out-of-sample period. The broad range of combinations allowed the construction of summary statistics for the RMSE obtained for all tested model types. These summary statistics are presented in Tables 7 and 8, categorized by model type and index. It is important to notice that both NAIVE and SEASONAL_NAIVE models are not shown in the tables, since there were no hyperparameters to select. The ENSEMBLE_BLENDING (MEAN) and ENSEMBLE_BLENDING (MEDIAN) models were constructed solely for the out-of-sample forecasting procedure.

The selected hyperparameters for each model type and index combination can be found in the Appendix.

The following conclusions can be drawn from the results presented in Tables 7 and 8:

(a) Either HOLT-WINTERS or TBATS models had the worst RMSE in validation, except for DURABLES index in which the TBATS model outperformed the SARIMA_LOG model;
(b) Calendar models, in general, surpassed their non-calendar counterparts;
(c) For all indices, XGBOOST was the best performing model when compared to other non-calendar models;
(d) The calendar version of one of the non-traditional models (MLP_CALENDAR, LSTM_CALENDAR, PROPHET_CALENDAR, PROPHET_ LOG_CALENDAR, and XGBOOST_CALENDAR) achieved the lowest validation RMSE across all indices;
(e) The widest RMSE range was attained by models SARIMA, SARIMA_LOG, SARIMA_CALENDAR, and SARIMA_LOG_CALENDAR. These findings were anticipated, and they indicate that the SARIMA model is sensible during parameter selection. Traditional SARIMA model selection approaches in the literature often involve evaluating one or more information criteria, such as Akaike and Schwarz, as seen in Pasquotto (2011), Angelo et al. (2011), and Bessa (2018).

Table 7
RMSE summary statistics for RETAIL and NONDURABLE indices by model type

index	model	validation RMSE statistics					
		min. (chosen)	Q1	median	Q3	max.	mean
RETAIL	HOLT-WINTERS	18.72	19.63	22.58	25.84	26.85	22.76
	TBATS	21.34	32.22	68.82	103.16	107.3	66.57
	SARIMA	16.08	17.28	18.02	18.71	171.9	25.93
	SARIMA_LOG	16.70	17.58	18.27	19.00	701.0	40.92
	MLP	13.91	14.52	14.99	15.93	28.25	15.56
	LSTM	13.64	14.04	14.24	14.51	18.35	14.47
	PROPHET	14.41	14.52	14.68	14.75	15.89	14.72
	XGBOOST	11.71	14.04	14.57	15.29	24.22	16.08
	ETS_CALENDAR	13.27	14.20	15.15	17.02	20.84	15.66
	ETS_LOG_CALENDAR	10.80	11.22	13.62	14.59	15.78	13.05
	SARIMA_CALENDAR	10.81	12.33	13.04	13.75	156.9	19.73
	SARIMA_LOG_CALENDAR	10.19	10.77	11.41	19.64	397.9	28.43
	MLP_CALENDAR	9.61	11.16	11.77	12.43	20.43	12.18
	LSTM_CALENDAR	9.91	10.41	10.67	11.06	16.03	10.89
	PROPHET_CALENDAR	10.33	10.53	10.70	11.08	15.01	11.19
	PROPHET_LOG_CALENDAR	9.79	10.12	10.28	10.43	11.03	10.31
	XGBOOST_CALENDAR	10.83	13.79	14.98	15.51	24.22	15.92
NONDURABLES	HOLT-WINTERS	30.78	31.02	31.18	31.49	32.08	31.28
	TBATS	64.53	72.11	92.61	121.33	153.6	100.8
	SARIMA	27.70	28.49	28.84	30.38	52.87	29.66
	SARIMA_LOG	27.16	28.35	29.21	31.40	66.33	30.87
	MLP	19.63	21.24	21.94	23.33	46.47	22.83
	LSTM	19.62	20.36	20.78	21.18	27.70	20.99
	PROPHET	21.72	22.31	23.47	23.95	25.95	23.37
	XGBOOST	18.44	20.88	21.87	25.48	38.81	24.55
	ETS_CALENDAR	21.45	23.73	24.88	27.82	52.20	26.26
	ETS_LOG_CALENDAR	16.12	16.80	22.88	23.85	25.11	21.09
	SARIMA_CALENDAR	19.01	24.70	49.75	63.37	72.76	44.84
	SARIMA_LOG_CALENDAR	15.75	19.44	22.52	36.06	240.7	32.46
	MLP_CALENDAR	14.95	16.26	17.21	18.29	41.46	17.81
	LSTM_CALENDAR	14.80	15.50	15.72	16.26	20.75	16.06
	PROPHET_CALENDAR	17.72	18.34	19.13	21.20	26.08	20.20
	PROPHET_LOG_CALENDAR	18.93	20.16	20.60	22.69	26.80	21.43
	XGBOOST_CALENDAR	15.98	20.40	21.53	25.05	38.81	23.78

Table 8
RMSE summary statistics for DURABLE and SERVICES indices by model type

index	model	validation RMSE statistics					
		min. (chosen)	Q1	median	Q3	max.	mean
	HOLT-WINTERS	53.36	53.68	58.69	63.66	63.93	58.69
	TBATS	40.42	41.13	44.11	90.79	222.6	87.81
	SARIMA	43.34	46.77	49.03	51.69	375.5	73.72
	SARIMA_LOG	46.11	46.85	47.67	53.30	35899	621.1
	MLP	25.56	26.81	27.65	28.84	39.81	28.08
	LSTM	25.17	26.38	27.09	27.58	37.33	27.12
	PROPHET	26.43	26.54	26.66	27.29	45.50	29.26
	XGBOOST	18.29	22.83	25.72	27.76	39.58	25.80
DURABLES	ETS_CALENDAR	24.65	32.21	40.54	55.63	1x10^12	2x10^10
	ETS_LOG_CALENDAR	23.06	24.75	28.88	39.39	42.80	31.07
	SARIMA_CALENDAR	18.69	19.75	22.68	26.64	278.6	39.51
	SARIMA_LOG_CALENDAR	17.97	18.93	20.23	31.52	1503	69.69
	MLP_CALENDAR	15.80	18.14	18.82	19.75	28.00	19.30
	LSTM_CALENDAR	16.27	17.13	17.59	18.66	22.20	17.95
	PROPHET_CALENDAR	15.95	16.75	17.00	26.24	39.64	20.54
	PROPHET_LOG_CALENDAR	15.08	15.53	15.73	16.30	24.22	16.22
	XGBOOST_CALENDAR	15.97	20.96	26.07	27.65	39.58	25.13
	HOLT-WINTERS	26.63	27.19	27.50	27.83	28.95	27.61
	TBATS	23.98	26.34	30.15	39.72	59.37	35.91
	SARIMA	20.36	20.57	21.64	25.72	257.5	34.09
	SARIMA_LOG	20.34	20.55	21.81	24.67	970.8	47.56
	MLP	13.34	14.75	15.42	16.74	32.41	16.09
	LSTM	13.59	14.07	14.38	14.85	19.17	14.70
	PROPHET	15.23	15.64	17.62	20.89	29.00	18.73
	XGBOOST	13.88	14.74	15.82	18.34	31.37	18.33
SERVICES	ETS_CALENDAR	16.12	17.68	18.71	20.48	26.81	19.56
	ETS_LOG_CALENDAR	14.52	15.44	16.99	18.62	21.78	17.36
	SARIMA_CALENDAR	13.83	14.68	19.33	27.33	244.2	31.68
	SARIMA_LOG_CALENDAR	14.06	14.79	15.41	18.75	721.9	50.84
	MLP_CALENDAR	12.73	13.49	13.97	14.78	21.80	14.38
	LSTM_CALENDAR	12.49	13.04	13.32	13.70	16.58	13.51
	PROPHET_CALENDAR	12.31	13.08	13.88	19.17	24.13	15.88
	PROPHET_LOG_CALENDAR	12.65	12.85	13.37	19.06	29.10	15.88
	XGBOOST_CALENDAR	13.19	14.75	15.36	17.77	30.64	18.15

4.2 Out-of-sample forecasting results

Out-of-sample forecasting, as depicted in Figure 3, was performed after the selection of the final hyperparameters for each index and model type. Table 9 summarizes the RMSE results for each index and time prediction combination. Results in bold indicate the minimum error obtained for each time horizon of a specific index.

For the model yielding the lowest RMSE in each time horizon, the daily forecasting squared error is illustrated in Figure 4. The charts are limited to 2000 for better visualization, due to outliers exceeding this value. Notably, error peaks seem to coincide across all indices, suggesting potential causes such as non-controlled calendar effects, unforeseen retail supply shocks—like Brazil's truck drivers' strike in May 2018—or data collection issues on specific days.

As outlined in the conclusion of Section 3, the results are analyzed from three distinct perspectives: non-calendar models, calendar models, and ensemble methods.

For the first analysis, evaluating non-calendar models (comparison for models #1 to #10 in Table 9), Table 10 summarizes the top-performing models for each index and time horizon combination:

By evaluating both Tables 9 and 10, the following insights can be derived:

(a) The XGBoost algorithm most frequently yielded the best RMSE in out-of-sample tests (8 out of 16 cases), followed by the Prophet model (6 out of 16 cases). This result corroborates findings by Bessa (2021), who also identified XGBoost as the top-performing model for the fashion retail industry;
(b) TBATS emerged as the preferred model for both NONDURABLES and SERVICES indices at the $D+1$ time prediction horizon. However, its RMSE performance deteriorates at longer prediction horizons. For the RETAIL and DURABLE indices, it was outpaced by a simple NAIVE model at $D+90$;
(c) NAIVE and SEASONAL NAIVE, in general, produced the worst RMSE results, across all indices. This result was expected due to the simplicity of these models. Usually these models serve as a baseline for gauging performance improvements when forecasting with more complex models;
(d) While artificial neural networks like MLP and LSTM occasionally outshone conventional time series forecasting models, they were significantly outpaced by machine learning algorithms like PROPHET and XGBOOST. This suggests the latter's superior predictive prowess, especially at extended time prediction horizons.

For the second analysis, which incorporates calendar models in the RMSE

Table 9
Out-of-sample RMSE results by index, time prediction horizon and model type

		out-of-sample RMSE results															
		RETAIL				NONDURABLES				DURABLES				SERVICES			
#	model	D+1	D+7	D+30	D+90	D+1	D+7	D+30	D+90	D+1	D+7	D+30	D+90	D+1	D+7	D+30	D+90
1	NAIVE	34.93	21.04	41.33	39.67	50.55	35.16	68.87	66.66	47.11	34.97	68.40	78.70	25.62	17.47	40.56	37.30
2	SEASONAL NAIVE	20.93	21.04	24.86	27.77	35.11	35.16	50.49	53.20	34.69	34.97	56.90	69.48	17.43	17.47	26.76	31.60
3	HOLT-WINTERS	16.25	18.53	21.12	23.61	26.65	32.06	44.41	51.38	24.94	31.75	52.03	67.15	14.41	16.65	24.44	30.44
4	TBATS	15.78	17.53	26.09	48.12	25.59	29.88	46.12	67.18	25.33	31.03	52.46	86.96	12.96	14.29	19.77	27.45
5	SARIMA	16.18	19.15	22.14	24.05	27.46	36.97	57.22	61.99	25.03	31.89	43.63	47.64	14.11	16.74	24.83	29.77
6	SARIMA_LOG	16.18	17.74	20.25	22.79	27.35	38.72	55.13	58.80	26.16	31.99	42.50	47.08	14.76	17.04	25.24	30.12
7	MLP	16.24	18.46	20.36	23.09	28.70	37.09	45.03	49.49	27.09	31.97	41.16	47.16	13.43	16.71	22.81	25.23
8	LSTM	15.73	18.05	19.94	22.34	26.47	32.84	44.84	49.16	25.27	29.95	40.73	47.99	12.98	15.89	22.42	25.22
9	PROPHET	15.27	15.85	16.16	17.20	25.64	28.19	29.42	33.63	24.14	25.14	25.65	27.69	13.59	14.25	14.92	16.28
10	XGBOOST	14.58	15.65	15.81	17.55	27.02	25.86	29.42	34.17	18.74	20.03	23.61	23.04	13.20	14.82	15.43	16.78
11	ETS_CALENDAR	13.50	13.65	14.07	14.85	25.67	26.33	27.63	31.09	20.85	21.05	21.76	23.26	14.02	14.36	15.42	17.38
12	ETS_LOG_CALENDAR	10.63	11.11	11.60	12.28	17.94	22.14	25.32	25.60	15.43	18.33	23.82	25.16	11.18	12.54	14.64	15.46
13	SARIMA_CALENDAR	9.96	10.99	11.66	13.84	18.54	21.58	25.39	28.87	13.83	15.22	18.21	19.42	10.90	11.84	13.94	16.01
14	SARIMA_LOG_CALENDAR	9.65	10.75	11.17	13.88	**16.42**	19.92	**20.40**	24.31	12.88	13.87	14.83	15.87	11.83	12.62	15.57	16.78
15	MLP_CALENDAR	11.21	13.54	13.92	13.71	20.77	23.94	28.07	30.83	16.30	19.84	19.49	19.98	11.23	13.31	14.48	15.08
16	LSTM_CALENDAR	9.85	10.53	10.92	11.95	17.52	20.15	24.13	27.32	13.40	16.11	18.31	18.93	10.43	11.77	12.64	14.24
17	PROPHET_CALENDAR	9.35	9.96	**10.16**	**10.71**	23.29	25.85	28.40	34.11	13.44	14.53	15.08	16.03	10.56	11.48	**11.97**	12.98
18	PROPHET_LOG_CALENDAR	9.51	10.39	10.47	10.84	19.98	22.87	23.08	25.55	12.06	13.92	14.15	15.71	10.10	11.42	12.13	**12.77**
19	XGBOOST_CALENDAR	12.07	12.02	12.33	13.16	23.79	24.63	28.78	32.13	15.34	15.03	15.20	17.37	12.99	13.54	14.74	15.24
20	ENSEMBLE_BLENDING (MEAN)	**9.14**	**9.82**	10.18	10.94	16.81	**19.35**	21.49	**23.14**	10.81	12.06	13.32	14.24	**9.92**	**10.88**	12.26	13.03
21	ENSEMBLE_BLENDING (MEDIAN)	9.15	9.91	10.33	10.91	16.86	19.77	22.61	24.18	**10.51**	**11.76**	**13.11**	**14.14**	10.22	11.25	12.50	13.20

Figure 4
Daily forecasting squared error plot for the minimum RMSE model by index and time prediction horizon

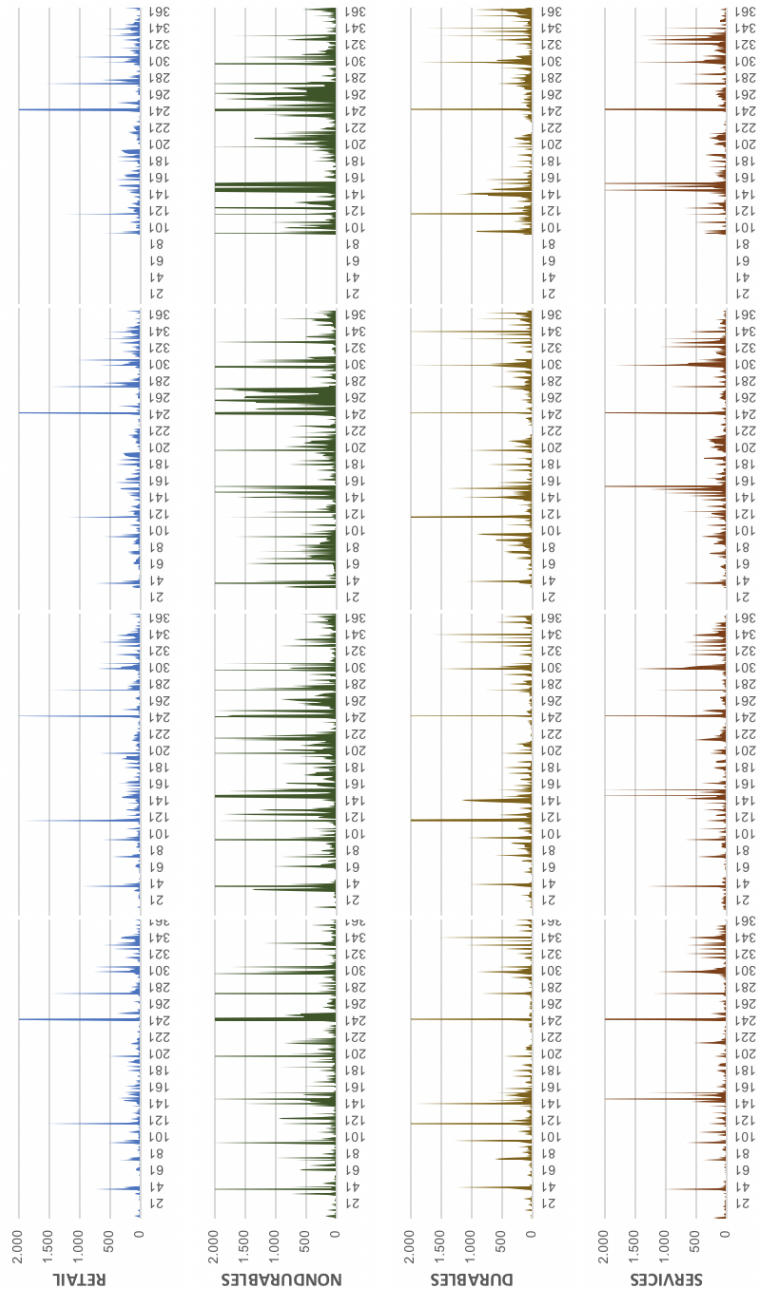


Table 10
Best performing non-calendar model (#1 to #10) by index and time horizon

index	horizon	best model
RETAIL	D+1	XGBOOST
	D+7	XGBOOST
	D+30	XGBOOST
	D+90	PROPHET
NONDURABLES	D+1	TBATS
	D+7	XGBOOST
	D+30	PROPHET
	D+90	PROPHET
DURABLES	D+1	XGBOOST
	D+7	XGBOOST
	D+30	XGBOOST
	D+90	XGBOOST
SERVICES	D+1	TBATS
	D+7	PROPHET
	D+30	PROPHET
	D+90	PROPHET

assessment (models #11 to #19 in Table 9), Table 11 presents evaluations of all individual models, summarizing the outcomes for each index and time horizon combination:

By evaluating both Tables 9 and 11, the following insights can be derived:

(a) For RETAIL, DURABLES, and SERVICES indices, PROPHET_CALENDAR or PROPHET_LOG_CALENDAR were predominantly the most accurate predictive models (11 out of 12 cases). For the NONDURABLE index, the model SARIMA_LOG_CALENDAR consistently achieved the best RMSE results across all time prediction horizons. These findings add to the ongoing debate on whether advanced models can surpass traditional time series ones;
(b) For NONDURABLE index, it is interesting to compare the performances of SARIMA_CALENDAR and SARIMA_LOG_CALENDAR. The log transformation proved crucial in boosting the predictive strength of the SARIMA calendar model. Without this transformation, it failed to outperform other sophisticated models, such as LSTM_CALENDAR;
(c) Surprisingly, with the inclusion of calendar variables, XGBOOST and XGBOOST_CALENDAR didn't yield the lowest RMSE scores. This outcome

Table 11
Best performing single model (#1 to #19) by index and time horizon

index	horizon	best model
RETAIL	D+1	PROPHET_CALENDAR
	D+7	PROPHET_CALENDAR
	D+30	PROPHET_CALENDAR
	D+90	PROPHET_CALENDAR
NONDURABLES	D+1	SARIMA_LOG_CALENDAR
	D+7	SARIMA_LOG_CALENDAR
	D+30	SARIMA_LOG_CALENDAR
	D+90	SARIMA_LOG_CALENDAR
DURABLES	D+1	PROPHET_LOG_CALENDAR
	D+7	SARIMA_LOG_CALENDAR
	D+30	PROPHET_LOG_CALENDAR
	D+90	PROPHET_LOG_CALENDAR
SERVICES	D+1	PROPHET_LOG_CALENDAR
	D+7	PROPHET_LOG_CALENDAR
	D+30	PROPHET_CALENDAR
	D+90	PROPHET_LOG_CALENDAR

contrasts with the observations of Bessa (2021);

(d) Except for the comparison of PROPHET and PROPHET_CALENDAR models at $D+90$ prediction horizon for the NONDURABLES index, all other model types benefited from the addition of calendar control variables. RMSE values improved markedly when comparing the calendar-enhanced and original versions of each model type. These results partly resonate with the work of Bessa (2021), who reported a slight improvement by adding seasonality treatment to fashion retail time series when forecasting with tree-based algorithm, such as the XGBoost. Two possible explanations for this divergence is that Bessa (2021) analyzed weekly data, which naturally tends to be less noisy than the daily indices evaluated in this text and the different seasonality treatment that was applied in both studies.

In out-of-sample period, 54 observations have one of the control variables exhibited in Table 3, corresponding to 14.8% of the entire period. To evaluate each model's performance in capturing calendar effects, these data points were isolated from the rest of the out-of-sample period. RMSE values were then computed, and the models were evaluated against each other. Table 12

showcases the results for all calendar models, index and time prediction horizons:

Table 12
RMSE results for calendar effect-controlled observations

model type	horizon	RETAIL	NONDURABLES	DURABLES	SERVICES
ETS_CALENDAR	D+1	24.09	42.06	39.48	22.19
ETS_LOG_CALENDAR	D+1	11.34	18.39	25.00	14.50
SARIMA_CALENDAR	D+1	10.96	21.02	17.88	14.05
SARIMA_LOG_CALENDAR	D+1	9.32	15.67	15.44	13.37
MLP_CALENDAR	D+1	11.74	25.74	22.64	14.73
LSTM_CALENDAR	D+1	9.82	16.75	17.67	11.72
PROPHET_CALENDAR	D+1	5.54	17.43	**10.65**	**7.89**
PROPHET_LOG_CALENDAR	D+1	**6.95**	**11.54**	13.73	8.76
XGBOOST_CALENDAR	D+1	18.54	30.31	29.17	17.32
ETS_CALENDAR	D+7	24.60	42.43	40.13	22.69
ETS_LOG_CALENDAR	D+7	12.21	20.85	29.30	15.57
SARIMA_CALENDAR	D+7	11.95	24.70	18.77	14.86
SARIMA_LOG_CALENDAR	D+7	9.17	17.86	15.25	14.49
MLP_CALENDAR	D+7	14.17	28.28	27.32	16.25
LSTM_CALENDAR	D+7	9.78	22.77	21.33	12.50
PROPHET_CALENDAR	D+7	**7.30**	23.45	**13.84**	**10.55**
PROPHET_LOG_CALENDAR	D+7	10.39	**16.15**	19.29	12.30
XGBOOST_CALENDAR	D+7	18.04	34.44	21.54	16.83
ETS_CALENDAR	D+30	25.92	44.37	42.02	24.11
ETS_LOG_CALENDAR	D+30	14.44	31.01	38.99	17.77
SARIMA_CALENDAR	D+30	13.46	30.52	21.55	16.73
SARIMA_LOG_CALENDAR	D+30	11.48	19.76	15.32	14.87
MLP_CALENDAR	D+30	16.12	32.62	28.21	17.15
LSTM_CALENDAR	D+30	11.32	30.93	23.83	15.60
PROPHET_CALENDAR	D+30	**7.64**	27.65	**14.64**	**11.28**
PROPHET_LOG_CALENDAR	D+30	9.95	**18.70**	20.06	12.67
XGBOOST_CALENDAR	D+30	18.17	42.77	20.80	19.52
ETS_CALENDAR	D+90	27.79	49.49	44.85	26.22
ETS_LOG_CALENDAR	D+90	14.75	38.71	39.59	18.64
SARIMA_CALENDAR	D+90	14.03	33.57	23.56	18.45
SARIMA_LOG_CALENDAR	D+90	13.93	31.02	18.02	18.65
MLP_CALENDAR	D+90	15.05	44.23	28.38	18.49
LSTM_CALENDAR	D+90	11.25	35.20	25.90	16.57
PROPHET_CALENDAR	D+90	**8.15**	34.00	**15.85**	**11.98**
PROPHET_LOG_CALENDAR	D+90	10.33	**25.93**	25.73	13.38
XGBOOST_CALENDAR	D+90	17.77	49.42	24.76	19.44

From Table 12, it is possible to verify that either PROPHET_CALENDAR or PROPHET_LOG_CALENDAR outperformed all the other models. Therefore, for forecasters prioritizing accuracy during specific holidays, both of the PROPHET calendar model variants seem ideal.

Finally, for the third analysis, both ensemble methods were incorporated into the model comparison (models #20 and #21 shown in Table 9). Table 13

now lists all models and summarizes the results:

Table 13
Best performing model (#1 to #21) by index and time horizon

index	horizon	best model
RETAIL	D+1	ENSEMBLE_BLENDING (MEAN)
	D+7	ENSEMBLE_BLENDING (MEAN)
	D+30	PROPHET_CALENDAR
	D+90	PROPHET_CALENDAR
NONDURABLES	D+1	SARIMA_LOG_CALENDAR
	D+7	ENSEMBLE_BLENDING (MEAN)
	D+30	SARIMA_LOG_CALENDAR
	D+90	ENSEMBLE_BLENDING (MEAN)
DURABLES	D+1	ENSEMBLE_BLENDING (MEDIAN)
	D+7	ENSEMBLE_BLENDING (MEDIAN)
	D+30	ENSEMBLE_BLENDING (MEDIAN)
	D+90	ENSEMBLE_BLENDING (MEDIAN)
SERVICES	D+1	ENSEMBLE_BLENDING (MEAN)
	D+7	ENSEMBLE_BLENDING (MEAN)
	D+30	PROPHET_CALENDAR
	D+90	PROPHET_LOG_CALENDAR

The findings in Table 13 echo those of Bessa (2021), in which an ensemble blending method can improve forecasting accuracy. Either by taking the mean or median of the forecasts generated by the calendar models—since the second analysis has shown they outperform their non-calendar counterpart—accuracy can be improved. For RETAIL, NONDURABLES and SERVICES indices ENSEMBLE_BLENDING (MEAN) was the best model in 6 out of 12 cases, while ENSEMBLE_BLENDING (MEDIAN) was the best model across all time prediction horizons for the DURABLES index.

A potential validation for the supremacy of these ensemble methods over individual models is the test proposed by Diebold and Mariano (1995) and modified by Harvey et al. (1997). The complete Diebold-Mariano test results are available in the Appendix. The primary takeaway is that ensemble methods aren't statistically distinguishable from the top-performing individual model presented in Table 11. As no single model consistently outperforms across all 16 index and time horizon combinations, adopting an ensemble approach is advised when forecasting Brazilian retail indices. This strategy

ensures results statistically on par with the best-performing model.

Conclusively, a comprehensive strategy for achieving superior results for all indices and time prediction horizons emerges: (1) Incorporate seasonality adjustments into the time series by examining holidays and calendar effects; (2) Make predictions for both the indices and their log transformations using conventional models, artificial neural networks, and machine learning algorithms; and, finally, (3) Ensemble all predictions, either averaging them or determining the median, to produce a singular forecast.

5. Final considerations

The primary objectives of this case study were: (1) to evaluate the existence of a single model capable of predicting daily indices with greater accuracy, comparing traditional models to artificial neural networks and machine learning algorithms; (2) to identify whether the results are consistent across the four sales indices; (3) to identify whether the results are consistent across prediction horizons; (4) to assess the impact on accuracy in modeling the calendar effect; (5) to identify whether an ensemble method produces superior forecasting results; and (6) to identify the feasibility of creating a consistent forecasting strategy for Brazilian retail sales data.

To ensure robustness in forecasts, the time series were divided into in-sample and out-of-sample periods. The in-sample period was further subdivided into a section for model estimation and another for validation. The split was performed through the holdout method, allocating 20% of the observations for the validation period. In total, 3372 different combinations of hyperparameters were tested for each index.

Through forecasting with an expanding window, the RMSE metric was calculated in the out-of-sample period across four different prediction horizons: $D+1$, $D+7$, $D+30$ and $D+90$. In response to the first proposed objective, it was not possible to identify a single model that performed better in all cases. PROPHET_CALENDAR, PROPHET_LOG_CALENDAR and SARIMA_LOG_CALENDAR emerged as the most accurate under varying conditions.

For objectives (2) and (3), it was possible to conclude that, for both the RETAIL and NONDURABLES indices, a single model delivered consistent performance. For the RETAIL index, PROPHET_CALENDAR outperformed all other individual models, while for the NONDURABLES index, SARIMA_LOG_CALENDAR was the winner. However, for the DURABLES and SERVICES indices, no single model outperformed the others across all prediction horizons, although PROPHET_LOG_CALENDAR was the most

recurring top-performer in both cases (leading in 6 out of 8 situations).

For objective (4), it was possible to draw a consistent conclusion that all model types benefited from integrating the calendar variables described in Section 3. By incorporating these variables, all calendar-based models outperformed their non-calendar counterparts, across all prediction horizons and indices, except for the PROPHET and PROPHET_CALENDAR model comparison at $D+90$ prediction horizon for the NONDURABLES index, which had a 1% RMSE increase when forecasting with calendar variables. This result underscores the great value in applying seasonality adjustments to daily time series indices. This finding partially aligns with Bessa (2021). A potential explanation could be the differing time frequencies in the two studies. Bessa (2021) assessed weekly time series, which are typically less volatile than daily data, emphasizing the importance of seasonality adjustments in the latter. Another possible reason is the distinct seasonal adjustment methods adopted in the two studies.

A Brazilian retail index forecaster's whose primary aim is to evaluate a specific holiday could benefit from employing the PROPHET_CALENDAR or PROPHET_LOG_CALENDAR models, which consistently report the lowest RMSE for calendar data points. An applicable scenario might be forecasting significant commercial events, such as Black Friday, which is increasingly influential in Brazilian retail, as highlighted by Guerra et al. (2017).

The inference from objective (5) is that ensemble methods enhance accuracy. ENSEMBLE_BLENDING (MEAN) or ENSEMBLE_BLENDING (MEDIAN) either yielded the lowest RMSE or were statistically on par with the top-performing model. This observation was consistent across all index and prediction horizon combinations. A fitting analogy for this observation is James Surowiecki's "Wisdom of Crowds" theory, which posits that a group's collective opinion often surpasses that of an individual expert. In this context, pooling information from a variety of calendar models can produce superior results compared to individual models.

Thus, the aforementioned findings provide insight into objective (6): it was feasible to devise a consistent forecasting strategy via the following steps: (1) applying seasonality adjustments to the time series by examining holidays and calendar effects; (2) generating predictions for the indices and their log transformations using traditional models, neural networks, and machine learning algorithms; and finally, (3) amalgamating them by either computing the mean or median prediction to produce a single forecast. It is pivotal to note that this strategy prioritizes accuracy enhancement. If considerations like computational time constrain the forecaster, this approach might not be ideal.

For instance, Bessa (2021) advocated for the seasonal Naïve model for fashion retail forecasting when the Analytic Hierarchy Process (AHP) method is utilized.

In conclusion, several factors might have restricted potential accuracy gains in the predictions. These include not incorporating exogenous variables from Brazil's macroeconomic landscape, not deploying intricate neural network designs (with a greater number of hidden layers), opting for the hold-out technique for hyperparameter selection, and not employing other seasonality adjustment and ensemble methodologies.

References

Almeida, F. C. d. and Passari, A. F. L. (2006). Previsão de vendas no varejo por meio de redes neurais, *Revista de Administração* **41**(3): 257–272.
URL: https://www.revistas.usp.br/rausp/article/view/44404

Alon, I., Qi, M. and Sadowski, R. J. (2001). Forecasting aggregate retail sales:: A comparison of artificial neural networks and traditional methods, *Journal of Retailing and Consumer Services* **8**(3): 147–156.
URL: https://doi.org/10.1016/S0969-6989(00)00011-4

Angelo, C. F. d., Zwicker, R., Fouto, N. M. M. D. and Luppe, M. R. (2011). Temporal series and neural networks: A comparative analysis of techniques in the Brazilian retail sales forecast, *Brazilian Business Review* **8**(2): 1–21.
URL: https://doi.org/10.15728/bbr.2011.8.2.1

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection, *Statistics Surveys* **4**: 40–79.
URL: https://doi.org/10.1214/09-SS054

Armstrong, J. S. and Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons, *International journal of forecasting* **8**(1): 69–80.
URL: https://doi.org/10.1016/0169-2070(92)90008-W

Au, K.-F., Choi, T.-M. and Yu, Y. (2008). Fashion retail forecasting by evolutionary neural networks, *International Journal of Production Economics* **114**(2): 615–630.
URL: https://doi.org/10.1016/j.ijpe.2007.06.013

Aye, G. C., Balcilar, M., Gupta, R. and Majumdar, A. (2015). Forecasting aggregate retail sales: The case of South Africa, *International Journal of Production Economics* **160**: 66–79.
URL: https://doi.org/10.1016/j.ijpe.2014.09.033

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation, *Information Sciences* **191**: 192–213.
URL: https://doi.org/10.1016/j.ins.2011.12.028

Bergmeir, C., Costantini, M. and Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation, *Computational Statistics & Data Analysis* **76**: 132–143.
URL: https://doi.org/10.1016/j.csda.2014.02.001

Bessa, A. B. (2018). *Previsão de vendas no varejo de moda com modelos de redes neurais*, Master's thesis, Fundação Getulio Vargas School of Economics.
URL: http://hdl.handle.net/10438/23890

Bessa, A. B. (2021). *Ensaios sobre previsão de vendas no varejo de moda*, PhD thesis, Fundação Getulio Vargas School of Economics.
URL: https://hdl.handle.net/10438/31137

Burman, P., Chow, E. and Nolan, D. (1994). A cross-validatory method for dependent data, *Biometrika* **81**(2): 351–358.
URL: https://doi.org/10.1093/biomet/81.2.351

Chatfield, C. (1988). Apples, oranges and mean square error, *International Journal of Forecasting* **4**(4): 515–518.
URL: https://doi.org/10.1016/0169-2070(88)90127-6

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, Association for Computing Machinery, New York, NY, USA, pp. 785–794.
URL: https://doi.org/10.1145/2939672.2939785

Chu, C.-W. and Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting, *International Journal of Production Economics* **86**(3): 217–231.
URL: https://doi.org/10.1016/S0925-5273(03)00068-9

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting, *International Journal of Forecasting* **22**(3): 443–473.
URL: https://doi.org/10.1016/j.ijforecast.2006.01.001

De Livera, A. M., Hyndman, R. J. and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing, *Journal of the American Statistical Association* **106**(496): 1513–1527.
URL: https://doi.org/10.1198/jasa.2011.tm09771

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business & Economic Statistics* **13**(3): 253–263.
URL: https://doi.org/10.1080/07350015.1995.10524599

Fang, W.-X., Lan, P.-C., Lin, W.-R., Chang, H.-C., Chang, H.-Y. and Wang, Y.-H. (2019). Combine Facebook Prophet and LSTM with BPNN forecasting financial markets: The Morgan Taiwan index, *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–2.
URL: https://doi.org/10.1109/ISPACS48206.2019.8986377

Frank, C., Garg, A., Sztandera, L. and Raheja, A. (2003). Forecasting women's apparel sales using mathematical modeling, *International Journal of Clothing Science and Technology* **15**(2): 107–125.
URL: https://doi.org/10.1108/09556220310470097

Garlapati, A., Krishna, D. R., Garlapati, K., Srikara Yaswanth, N. m., Rahul, U. and Narayanan, G. (2021). Stock price prediction using Facebook Prophet and Arima models, *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–7.

Gers, F. A., Schmidhuber, J. and Cummins, F. (2000). Learning to forget: Continual prediction with LSTM, *Neural Computation* **12**(10): 2451–2471.

Geurts, M. D. and Kelly, J. P. (1986). Forecasting retail sales using alternative models, *International Journal of Forecasting* **2**(3): 261–272.
URL: https://doi.org/10.1016/0169-2070(86)90046-4

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks, *in* Y. W. Teh and M. Titterington (eds),

Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Vol. 9 of *Proceedings of Machine Learning Research*, PMLR, Chia Laguna Resort, Sardinia, Italy, pp. 249–256.
URL: https://proceedings.mlr.press/v9/glorot10a.html

Glorot, X., Bordes, A. and Bengio, Y. (2011). Deep sparse rectifier neural networks, *in* G. Gordon, D. Dunson and M. Dudík (eds), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Vol. 15 of *Proceedings of Machine Learning Research*, PMLR, Fort Lauderdale, FL, USA, pp. 315–323.
URL: https://proceedings.mlr.press/v15/glorot11a.html

Gorr, W. L. (1994). Editorial: Research prospective on neural network forecasting, *International Journal of Forecasting* **10**(1): 1–4.
URL: https://doi.org/10.1016/0169-2070(94)90044-2

Guerra, A., Nielsen, F. and Ghisi, M. (2017). Aspectos benéficos e detratores do Black Friday no Brasil: Um estudo sobre as práticas adotadas pelos varejistas para essa data promocional, *14º Congresso Latino-Americano de Varejo e Consumo*, CLAV, São Paulo.
URL: http://bibliotecadigital.fgv.br/ocs/index.php/clav/clav2017/paper/viewPaper/6021

Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors, *International Journal of Forecasting* **13**(2): 281–291.
URL: https://doi.org/10.1016/S0169-2070(96)00719-4

Hill, T., Marquez, L., O'Connor, M. and Remus, W. (1994). Artificial neural network models for forecasting and decision making, *International Journal of Forecasting* **10**(1): 5–15.
URL: https://doi.org/10.1016/0169-2070(94)90045-0

Hippert, H., Bunn, D. and Souza, R. (2005). Large neural networks for electricity load forecasting: Are they overfitted?, *International Journal of Forecasting* **21**(3): 425–434.
URL: https://doi.org/10.1016/j.ijforecast.2004.12.004

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation* **9**(8): 1735–1780.

Hyndman, R. J., Koehler, A. B., Snyder, R. D. and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting* **18**(3): 439–454.
URL: https://doi.org/10.1016/S0169-2070(01)00110-8

Hyndman, R., Koehler, A., Ord, K. and Snyder, R. (2008). *Forecasting with exponential smoothing: The state space approach*, Springer Series in Statistics, Springer Berlin, Heidelberg.
URL: https://doi.org/10.1007/978-3-540-71918-2

Ji, S., Wang, X., Zhao, W. and Guo, D. (2019). An application of a three-stage XGBoost-based model to sales forecasting of a cross-border e-commerce enterprise, *Mathematical Problems in Engineering* **2019**: 8503252.
URL: https://doi.org/10.1155/2019/8503252

Karlik, B. and Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks, *International Journal of Artificial Intelligence and Expert Systems* **1**(4): 111–122.
URL: https://www.cscjournals.org/library/manuscriptinfo.php?mc=IJAE-26

Lawrence, M., O'Connor, M. and Edmundson, B. (2000). A field study of sales forecasting accuracy and processes, *European Journal of Operational Research* **122**(1): 151–160.
URL: https://doi.org/10.1016/S0377-2217(99)00085-5

Maas, A. L., Hannun, A. Y. and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models, *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, International Conference on Machine Learning.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition, *Journal of Forecasting* **1**(2): 111–153.
URL: https://doi.org/10.1002/for.3980010202

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K. and Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study, *International Journal of Forecasting* **9**(1): 5–22.
URL: https://doi.org/10.1016/0169-2070(93)90044-N

Makridakis, S. G., Wheelwright, S. C. and Hyndman, R. J. (1998). *Forecasting: Methods and applications*, 3rd edn, Wiley, New York.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines, *Proceedings of the 27th ICML*, International Conference on Machine Learning, pp. 807–814.
URL: https://icml.cc/Conferences/2010/papers/432.pdf

Pasquotto, J. L. D. (2011). *Previsão de séries temporais no varejo brasileiro: Uma investigação comparativa da aplicação de redes neurais recorrentes de elman*, Master's thesis, School of Economics, Business and Accounting – Universidade de São Paulo.
URL: https://doi.org/10.11606/D.12.2011.tde-24022011-180352

Pinto, J. M. and Marçal, E. F. (2021). Cross-validation based forecasting method: A machine learning approach, *SSRN Working Paper 3326496*, Social Science Research Network.
URL: http://doi.org/10.2139/ssrn.3326496

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation, *Journal of Econometrics* **99**(1): 39–61.
URL: https://doi.org/10.1016/S0304-4076(00)00030-0

Ruder, S. (2016). An overview of gradient descent optimization algorithms.
URL: https://doi.org/10.48550/arXiv.1609.04747

Saradhi, R. P. and Naresh, N. (2018). An advanced sales forecasting system using XGBoost algorithm, *International Journal of Electrical Electronics Computers & Mechanical Engineering* **8**(7).

Shao, J. (1993). Linear model selection by cross-validation, *Journal of the American Statistical Association* **88**(422): 486–494.
URL: https://doi.org/10.1080/01621459.1993.10476299

Swanson, N. R. and White, H. (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models, *International Journal of Forecasting* **13**(4): 439–461.
URL: https://doi.org/10.1016/S0169-2070(97)00030-7

Taylor, S. J. and Letham, B. (2018). Forecasting at scale, *The American Statistician* **72**(1): 37–45.

Toharudin, T., Pontoh, R. S., Caraka, R. E., Zahroh, S., Lee, Y. and Chen, R. C. (2021). Employing long short-term memory and Facebook Prophet model in air temperature forecasting, *Communications in Statistics – Simulation and Computation* **0**(0): 1–24.

Zadeh, M. R., Amin, S., Khalili, D. and Singh, V. P. (2010). Daily outflow prediction by multi layer perceptron with logistic sigmoid and tangent sigmoid activation functions, *Water Resources Management* **24**(11): 2673–2688.
URL: https://doi.org/10.1007/s11269-009-9573-4

Zheng, W. (2019). hv-block cross validation is not a BIBD: A note on the paper by Jeff Racine.
URL: https://doi.org/10.48550/arxiv.1910.08904

A. Additional tables and figures

Figure A1
Summary statistics, autocorrelation and partial autocorrelation function plots by index

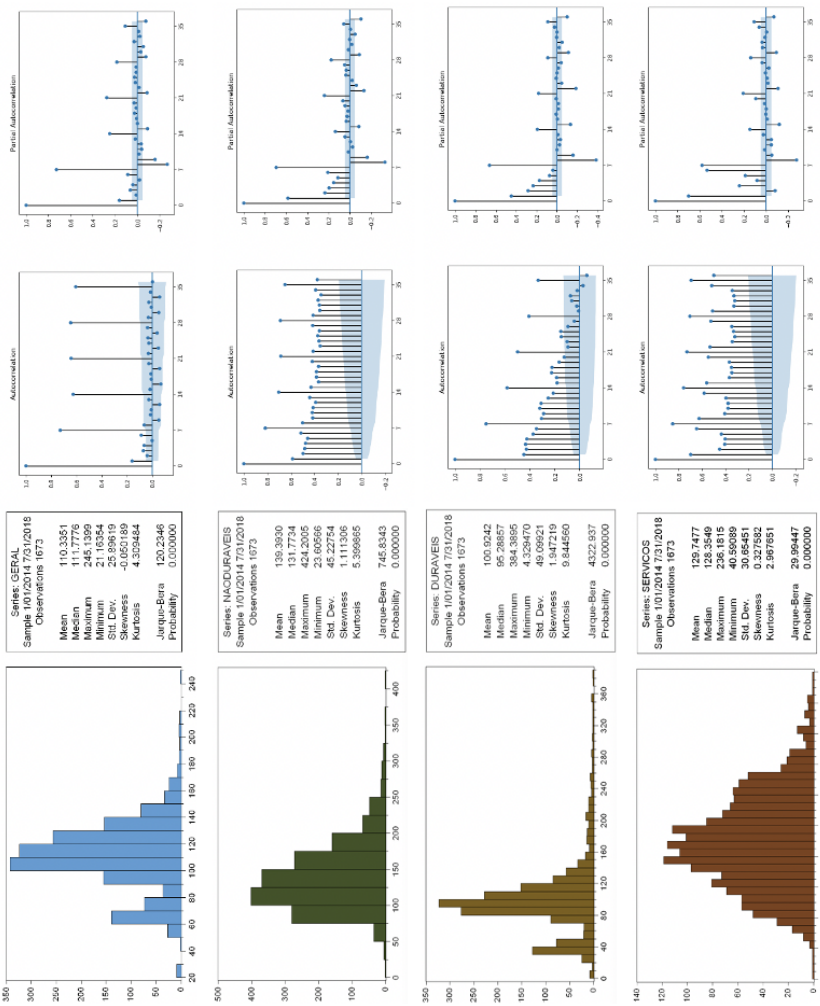


Table A1
Hyperparameters select by model and index in hold-out procedure: non-calendar models

#	model	hyperparameter	index			
			RETAIL	NONDURABLES	DURABLES	SERVICES
1	NAIVE					
2	SEASONAL NAIVE					
3	HOLT-WINTERS	seasonal cycle	35	28	14	28
		seasonal component	additive	additive	additive	additive
4	TBATS	box-cox transformation	yes	no	yes	no
		damped trend	yes	yes	no	yes
5	SARIMA	(p,d,q,P,D,Q,S)	(5,0,1,5,0,1,7)	(1,1,0,5,0,0,7)	(6,0,1,4,1,1,7)	(6,0,1,3,1,0,7)
6	SARIMA_LOG	(p,d,q,P,D,Q,S)	(6,1,0,5,0,0,7)	(1,1,0,5,0,0,7)	(4,0,1,5,1,1,7)	(6,0,1,3,1,0,7)
7	MLP	number of lags	14	28	7	7
		neurons in hidden layer	35	35	21	35
		learning rate	0.1	0.1	0.1	0.01
		activation function	sigmoid	ReLU	ReLU	ReLU
		optimizer	Adam	RMSprop	Adam	RMSprop
8	LSTM	number of lags	14	28	14	14
		neurons in hidden layer	14	35	7	21
		learning rate	0.1	0.01	0.01	0.1
		optimizer	Adam	RMSprop	Adam	Adam
9	PROPHET	changepoint scale	0.01	0.5	0.1	0.1
		seasonality scale	0.1	0.1	0.01	0.01
		seasonality mode	multiplicative	multiplicative	additive	multiplicative
10	XGBOOST	learning rate	0.06	0.1	0.04	0.08
		max depth	4	2	10	10
		min child weight	1	1	1	10
		gamma	100	0	10	100
		subsample	1	0.5	1	0.5

Table A2
Hyperparameters select by model and index in hold-out procedure: calendar models

#	model	hyperparameter	index			
			RETAIL	NONDURABLES	DURABLES	SERVICES
11	ETS_CALENDAR	seasonal component	(7,35,364)	(7,35,364)	(7,35,364)	(7,35,364)
		ETS structure	MMM	MAM	MAdA	MMM
12	ETS_LOG_CALENDAR	seasonal component	-7,35	-7,35	(7,35,364)	-7,35
		ETS structure	MAM	MAdA	MAdA	MAdA
13	SARIMA_CALENDAR	(p,d,q,P,D,Q,S)	(1,1,1,3,1,1,7)	(1,0,0,2,1,1,7)	(3,0,1,1,1,1,7)	(2,0,1,1,1,1,7)
14	SARIMA_LOG_CALENDAR	(p,d,q,P,D,Q,S)	(5,1,1,3,0,0,7)	(5,1,1,2,1,1,7)	(6,0,1,5,1,1,7)	(1,0,0,1,1,0,7)
15	MLP_CALENDAR	number of lags	7	21	21	7
		neurons in hidden layer	14	7	35	7
		learning rate	0.1	0.1	0.1	0.001
		activation function	ReLU	ReLU	ReLU	sigmoid
		optimizer	Adam	RMSprop	RMSprop	RMSprop
16	LSTM_CALENDAR	number of lags	35	35	28	35
		neurons in hidden layer	21	28	7	28
		learning rate	0.1	0.1	0.1	0.1
		optimizer	Adam	Adam	Adam	Adam
17	PROPHET_CALENDAR	changepoint scale	0.01	0.1	0.01	0.1
		seasonality scale	10	10	0.01	1
		seasonality mode	additive	additive	multiplicative	additive
18	PROPHET_LOG_CALENDAR	changepoint scale	0.001	0.5	0.1	0.5
		seasonality scale	10	10	0.01	0.01
		seasonality mode	additive	additive	multiplicative	additive
19	XGBOOST_CALENDAR	learning rate	0.1	0.1	0.08	0.1
		max depth	8	4	10	2
		min child weight	1	1	1	1
		gamma	100	100	10	0
		subsample	1	0.5	1	1

Table A3
RMSE reduction by model type, index and prediction horizon

index	model type	RMSE Reduction* D+1	D+7	D+30	D+90
RETAIL	ETS	−17%	−26%	−33%	−37%
	SARIMA	−38%	−43%	−47%	−42%
	SARIMA_LOG	−40%	−39%	−45%	−39%
	MLP	−31%	−27%	−32%	−41%
	LSTM	−37%	−42%	−45%	−47%
	PROPHET	−39%	−37%	−37%	−38%
	XGBOOST	−17%	−23%	−22%	−25%
NONDURABLES	ETS	−4%	−18%	−38%	−39%
	SARIMA	−32%	−42%	−56%	−53%
	SARIMA_LOG	−40%	−49%	−63%	−59%
	MLP	−28%	−35%	−38%	−38%
	LSTM	−34%	−39%	−46%	−44%
	PROPHET	−9%	−8%	−3%	1%
	XGBOOST	−12%	−5%	−2%	−6%
DURABLES	ETS	−16%	−34%	−58%	−65%
	SARIMA	−45%	−52%	−58%	−59%
	SARIMA_LOG	−51%	−57%	−65%	−66%
	MLP	−40%	−38%	−53%	−58%
	LSTM	−47%	−46%	−55%	−61%
	PROPHET	−44%	−42%	−41%	−42%
	XGBOOST	−18%	−25%	−36%	−25%
SERVICES	ETS	−3%	−14%	−37%	−43%
	SARIMA	−23%	−29%	−44%	−46%
	SARIMA_LOG	−20%	−26%	−38%	−44%
	MLP	−16%	−20%	−37%	−40%
	LSTM	−20%	−26%	−44%	−44%
	PROPHET	−22%	−19%	−20%	−20%
	XGBOOST	−2%	−9%	−4%	−9%

(*): Percentage RMSE variation of calendar effect-controlled model in comparison to its non-controlled version.

Table A4
Diebold-Mariano test results – best performing model vs. other model types at 5% significance level

	p-value															
	RETAIL				NONDURABLES				DURABLES				SERVICES			
model	D+1	D+7	D+30	D+90	D+1	D+7	D+30	D+90	D+1	D+7	D+30	D+90	D+1	D+7	D+30	D+90
NAIVE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SEASONAL NAIVE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HOLT-WINTERS	0.007	0.000	0.000	0.000	0.024	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TBATS	0.003	0.000	0.000	0.000	0.009	0.000	0.000	0.000	0.009	0.000	0.000	0.001	0.000	0.002	0.000	0.000
SARIMA	0.005	0.000	0.000	0.000	0.026	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SARIMA_LOG	0.001	0.000	0.000	0.000	0.025	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MLP	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.001	0.000	0.000	0.000
LSTM	0.004	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.041	0.000	0.000	0.000	0.002	0.000	0.000	0.000
PROPHET	0.001	0.001	0.002	0.003	0.002	0.000	0.002	0.000	0.004	0.004	0.008	0.012	0.002	0.003	0.011	0.010
XGBOOST	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.040	**0.060**	0.032	0.017	0.000	0.014	0.002	0.010
ETS_CALENDAR	0.000	0.000	0.001	0.002	0.000	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.000	0.001	0.002	0.000
ETS_LOG_CALENDAR	0.001	0.001	0.014	0.019	0.011	0.000	0.000	0.036	0.002	0.001	0.000	0.000	0.002	0.000	0.000	0.000
SARIMA_CALENDAR	0.017	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.060**	0.038	0.003	0.000
SARIMA_LOG_CALENDAR	**0.084**	0.000	0.014	0.000		**0.296**		**0.223**	0.010	0.008	0.045	0.044	0.001	0.002	0.000	0.000
MLP_CALENDAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.000	0.004	0.021
LSTM_CALENDAR	0.004	0.024	**0.068**	0.002	**0.129**	**0.164**	0.005	0.000	0.000	0.000	0.000	0.000	**0.084**	0.003	**0.352**	**0.124**
PROPHET_CALENDAR	**0.436**	**0.515**			0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.008	**0.250**	**0.197**		**0.509**
PROPHET_LOG_CALENDAR	**0.232**	**0.159**	**0.191**	**0.605**	0.002	0.000	0.011	0.044	**0.072**	**0.099**	**0.430**	**0.396**	**0.733**	**0.305**	**0.654**	
XGBOOST_CALENDAR	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	**0.141**	0.004	0.026	0.000	0.000	0.000	0.000	0.000
ENSEMBLE_BLENDING (MEAN)			**0.921**	**0.396**	**0.537**		**0.183**		**0.262**	**0.170**	**0.435**	**0.731**			**0.473**	**0.628**
ENSEMBLE_BLENDING (MEDIAN)	**0.917**	**0.429**	**0.560**	**0.492**	**0.472**	**0.092**	**0.063**	**0.050**					**0.213**	**0.069**	**0.263**	**0.433**