Retail sales forecasting for a Brazilian supermarket chain: an empirical assessment

Fernanda M. de Almeida
Supermercados Nordestão / UFRN
Natal, RN, Brazil
fernanda.almeida@nordestao.com.br

Allan M. Martins
DEE, UFRN
Natal, RN, Brazil
allan@dee.ufrn.br

Marcus A. Nunes
EST, UFRN
Natal, RN, Brazil
marcus.nunes@ufrn.br

Leonardo C. T. Bezerra
IMD, UFRN
Natal, RN, Brazil
leobezerra@imd.ufrn.br

***Abstract*—Time series forecasting is a consolidated, broadly used approach in several fields, such as finance and industry. Retail can also benefit from forecasting in many areas such as stock demand, price optimization, and sales. This study addresses retail sales forecasting in Nordestão, a large Brazilian supermarket chain that respectively ranks 3rd and 27th in sales regionally and nationally. The data considered spans five years of daily transactions from eight different stores. Knowingly effective machine learning techniques for forecasting are adopted, namely linear regression, random forests, and XGBoost. We further improve their performance with features we engineer to address seasonal effects. The best algorithm varies per store, but for most stores at least one of the methods proves effective. Importantly, the models display effective performance across multiple testing weeks, and improve over the current approach of Nordestão by a significant margin. Besides the traditional relevance of sales forecasting, our work is a means for Nordestão to evaluate the impact of the COVID-19 pandemics on sales.**

***Index Terms*—Time series forecasting, machine learning, supermarket retail sales**

I. Introduction

Data-driven decision-making has increasingly become central in the industry, ranging from strategic to tactical and operational decisions [1]. In retail, for instance, all organizations make *strategic* decisions that define their market approaches, and identify competitive factors that will guide technological developments and regulatory environments. In turn, defining appropriate promotional tools at the *tactical* level maximizes the overall profit of the chain and its stores. Strategic and tactical decisions are supported by the *operational* level, e.g., managing logistics, inventory, and distribution.

An effective data-driven decision-making culture requires prediction or forecasting, which is enabled by data availability. In particular, time series (TS) forecasting is a well-known approach adopted in the industry for anticipating demand [2]. From an algorithmic perspective, TS forecasting is a well-developed field, ranging from assumption-based to assumption-free algorithms [3]. The former includes linear regression (LR) and the AR(I)MA model family [4]. Among the latter, ensemble machine learning approaches such as random forests (RF [3]) and XGBoost (XGB [5]) have become methods of choice. Depending on data and computational infrastructure availability, *deep learning* approaches [6] may also be employed. The richness in options is due to the heuristic nature of the algorithms, which require that multiple alternatives be considered when facing a particular problem.

The goal of this study is to forecast retail sales from the Nordestão supermarket chain. The chain comprises nine retail and two wholesale stores in the metropolitan area of Natal, in the Northeast region of Brazil. Nordestão currently ranks 3rd in sales in the Northeast region, and 27th in sales nation-wide [7]. Recently, the COVID-19 pandemic accelerated the digital transformation of the company, and Nordestão expanded its digital channel sales. As the chain expands, its operations are becoming increasingly more complex. In this context, modeling the sales demand from Nordestão for tactical decision-making is seminal to address operational issues, such as stock planning and distribution.

In this work, we focus on retail stores that have not been recently inaugurated, and hence have enough available data for forecasting. For each of these stores, we use daily sales values from 2015 to 2019 to predict 2020 pre-pandemic values. Our rationale is that (i) tactical-level forecasting usually targets store-wise total sales [1], and (ii) the pandemic strongly affected sales from March 2020 on, and hence an initial modeling should focus on data prior to COVID-19. We enrich this data with (i) weekday and month information, encoded using trigonometric representations; (ii) trend and seasonality components [8], and; (iii) lagged versions of the raw data.

Given the data and infrastructure available, we investigate LR, RF, and XGB as forecasting algorithms.[1] Our experiments (i) evaluate the benefits of the features we engineer; (ii) compare the effectiveness of the algorithms selected, and; (iii) assess their improvement over baseline models, including the current approach from Nordestão. Regarding (i), we use walk-forward cross-validation on the 2015–2019 data, and confirm the importance of data enrichment. Nonetheless, we observe that lagged values and monthly information contribute the least among the features engineered, and could hence be removed for model simplicity.

Concerning (ii) and (iii), we train on 2015–2019 data and holdout 2020 data for testing, using two variability levels. At a lower level, we evaluate a model based on a 7-day prediction window, where to predict a given day the model is given access

[1]Tests with seasonal AR(I)MA were also performed, but the results obtained were below expectations, and therefore are not described in this work.

2022 IEEE 24th Conference on Business Informatics (CBI) | 978-1-6654-6016-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CBI54897.2022.00014

2378-1971/22/$31.00 ©2022 IEEE
DOI 10.1109/CBI54897.2022.00014

to real data from previous days. Our rationale is that evaluating models based on 7-day R^2 scores reduces weekday effects. At a higher level, we evaluate models based on their overall performance across the different stores and weeks, and observe how model performance varies as a function of these two factors. In detail, we consider the 10 pre-pandemic weeks in 2020. With the exception of the weeks affected by holidays or the announcement of the first social distancing measures, best-performing models achieve $R^2 > 0.8$ and median cross-store $R^2 > 0.6$. More importantly, models consistently outperform the baseline approaches considered.

The remainder of this paper is structured as follows. In Section II, we briefly define the time series problem we address, and discuss related work. The sections that follow are structured according to the CRISP-DM methodology [9]. In detail, Section III details the Nordestão supermarket chain business and data. In Section IV, we propose our forecasting pipeline, delimiting prediction algorithms, data preparation, and the experimental setup we adopt. We evaluate models in Sections V and VI, and conclude in Section VII highlighting future work opportunities and key findings that are reusable by other companies seeking to model their retail sales.

II. Background

In the context of retail sales forecasting, both time series analysis and machine learning play a critical role. In this section, we briefly define the forecasting problem we address. Later, we provide a broader discussion on the works that apply machine learning to retail sales forecasting.

A. Retail sales forecasting as a time series problem

A time series is a set of sequential data sampled in a specific time unit, used to record a process output to enable the analysis of its evolution [8]. In retail, data granularity varies as a function of decision-making level [1]. Strategic decisions are supported by total chain sales forecasting, sometimes aggregated regionally. For the tactical level, as conducted in this work, per-store demand forecasting is necessary, as demand may present regional seasonal effects. Last, at an operational level data granularity is usually defined at a store-level *stock keeping unit*.

A few components stand out in time series and can be used to model processes, namely trend (T), seasonality (S), and residuals (R). These three components formally comprise a time series $Y(t) = T(t) + S(t) + R(t)$. The residuals are the part of the data that has random behavior. In turn, trend is defined as the long-term effect of rising or falling, whereas seasonality is a pattern that repeats in a well-defined, frequent period of time. Different approaches have been proposed in the literature regarding trend and seasonality. Whichever the technique employed, a proper seasonal adjustment requires an investigation of seasonal patterns, which can often affect multiple periods (e.g. weekly and monthly). We next briefly discuss the methods that are most directly related to our work.

Moving average (MA) filters are an optimal linear strategy to reduce noise in time-encoded signals while preserving sharp steps [4]. Effectively, an MA comprises a series of consecutive time period averages. As such, MAs are defined as a function of (i) the number of days n that the period comprises and (ii) whether the period is centered or shifted. Another important aspect of MA as an estimator concerns periodic datasets. Taking advantage of the inherit periodicity, an MA can estimate using past data that correspond to the period index. If the data is indexed by days, for instance, there might be some weekday periodicity. In this case, one might use weekday-specific MAs as estimators.

STL [8] uses a nonparametric function to fit a smooth curve through weighted regression. The technique is able to isolate all components in a time series, namely trend, seasonality, and residuals. In comparison to MAs, STL is more robust as it presents a better treatment for outliers. In addition, STL deals with different types of seasonality in a configurable way.

B. Related work

The literature on retail forecasting is fragmented, most often focusing on specific segments. For instance, fashion retail has received significant attention, and several surveys on the topic can be identified [6]. Even the research on grocery retail is not limited to supermarkets, but also addresses specific niches such as convenience stores [10] and fruit supermarkets [11]. Grocery retail works may also specialize in food categories, as coarse as fresh food [10] or fine-grained as grapes [11]. Alternative formulations to retail sales forecasting concern item demand [2] and consumer-wise sale forecasting [12]. Demand and sale forecasting can also be embedded in supply management research [13]. Additional challenges are (i) scale [2], especially in online commerce scenarios, and; (ii) the impact of exogenous data, such as meteorological [14] and macroeconomical conditions [15].

Algorithm-wise, forecasting works generally adopt (i) traditional statistical models [4]; (ii) computational intelligence approaches either based on machine or deep learning [6], and; (iii) combinations of the previous categories using ensembles [13]. Yet, we were not able to identify large-scale, rigorous studies assessing the best approaches for supermarket retail in general. This is likely explained by the nature of these works, which traditionally address case studies rather than the forecasting problem from a broad perspective. The most closely related survey we identified discusses data preparation, modeling, and evaluation, but is limited as to algorithmic options and experimental discussion [16].

Finally, there are not many papers forecasting retail data in the Brazilian context. [17], for example, showed that neural networks have better results than naive techniques. On the other hand, [18] state that there is no difference between neural networks or ARIMA models when the sum of squares is considered as the metric for evaluation. When modeling the sales of a retail clothing company, [19] showed that a static forecast model is adequate for the company demand, but the small sample size could lead to estimation distortions.

As discussed in this section, retail sales forecasting builds on time series analysis, but the existing literature is fragmented

TABLE I: Retail stores considered in this work and their attributes.

ID	Neighborhood	Area (%)	Mix	Value (%)	p-value
5	Commercial	117	9618	6.76	$7.24\text{e}{-10}$
8	Residential	100	9795	6.80	$2.97\text{e}{-7}$
9	Commercial	276	16705	18.76	$1.71\text{e}{-10}$
12	Commercial	225	16282	16.23	$8.00\text{e}{-4}$
13	Residential	203	10721	11.97	$5.19\text{e}{-5}$
16	Residential	121	9302	7.59	$1.06\text{e}{-2}$
31	Residential	257	16317	16.60	$4.11\text{e}{-10}$
34	Residential	234	16281	13.11	$8.33\text{e}{-8}$
38	Commercial	214	16320	2.14	$1.10\text{e}{-5}$

into case studies. Over the next sections, we describe the approach we propose for Nordestão, grouped by the steps of the CRISP-DM methodology [9].

III. Business and Data Understading

The initial steps for an industrial data mining application is to understand the target business and its available data [9]. In this section, we first present the Nordestão supermarket chain stores, describing their most relevant characteristics. Later, we describe data acquisition and conduct an exploratory analysis of the data available at the company.

A. Business understanding

The Nordestão supermarket chain built a total of 12 stores during its lifetime, of which nine are retail stores and three are wholesale stores. Among these 12 stores, only one wholesale store is located outside the metropolitan area of Natal. Given the expected differences between retail and wholesale, we delimit our further discussion to retail stores, the focus of this paper. We remark that when this work started, one retail store had recently been opened. As such, data from this store is expected to be too few for proper forecasting, which we illustrate for completeness but do not model.

Table I provides an overview of the retail stores of the chain, which we label with integer identifiers for confidentiality. For each store, Table I gives: (i) the type of the neighborhood in which the store is located; (ii) the area ratio in relation to the smallest store; (iii) product mix, i.e. the number of distinct products available for purchase at the store, and; (iv) the sales participation value of the store w.r.t. to chain total considering the period between 2015 and 2019. We remark that (i) the p-value will later be discussed, and; (ii) we exclude from this overview sales data from 2020 on due to the effects of the COVID-19 pandemic.

From Table I, we notice that the retail stores of the chain are nearly evenly distributed among commercial and residential areas. More importantly, we observe that store area is strongly correlated to participation in sales, as follows. Stores 9 and 31, which are the largest stores, also present the highest sales share. The opposite is also true, with the smallest stores (5, 8, and 16) presenting the lowest share, except for the recently inaugurated store 38. For the remaining medium-sized stores, no clear pattern can be observed, though the neighborhood appears to play a role, as follows. Store 12 presents a very high participation value, being located in a commercial area. Conversely, stores 13 and 34 present moderate participation values, being located in residential areas.

B. Data understanding

Sales data is taken from the sales coupon database. As previously discussed, retail tactical-level forecasting is usually conducted from a per-store total daily sales perspective [1]. This is the approach we adopt in this work, which is further justified by the recent digital transformation of the company, as follows. Around 2017, the company changed its enterprise resource planning (ERP) software, and so its product-hierarchy assignment was altered. As such, merging data originated from different systems would inevitably lead to information loss. Next, we illustrate and discuss total sales for each retail store. Later, we investigate series stationarity, trend and seasonality for varying time periods, and probability distribution.

1) Daily sales and cross-store patterns: The per-store total daily sales is given in Figure 1 in blue, where the stores are labeled with the identifiers given in Table I. Values on the y-axis have been globally scaled to the $[0, 1]$ range for confidentiality. As previously discussed, we delimit the data assessed in this work to sales records from January 2015 to December 2019, and remark that data from store 38 is not enough for proper modeling. Importantly, due to the geographical proximity between stores 12 and 38, we observe a structural break in sales at the former when the latter is inaugurated. As such, we expect forecasting for store 12 to be less effective than for the remaining stores.

The global scaling we employ promotes confidentiality while preserving comparability between stores. For instance, we see how total sales from the smallest stores (5, 8 and 16) present ranges much lower in the y-axis than the remaining stores, with a smaller amplitude as well. In addition, we observe how the stores with largest participation values not only present ranges higher in the y-axis, but their peaks are much higher as well. In particular, we remark that store 12 could likely be among the highest participation ratios if not for the structural break starting on late 2018. Regarding strong peaks, for some stores these can be detrimental to forecasting, since peaks are considered outliers. From a business perspective, though, these peaks comprise large promotional events and holidays such as the December holiday shopping season. As such, forecasting these events are a challenge that should be addressed in future work.

2) Stationarity: Since assumption-based forecasting methods assume stationary series, we use the augmented Dickey-Fuller (ADF) test to verify this property. For the ADF test, the null hypothesis is that an autoregressive series contains a unit root. The alternative hypothesis is that the series is either stationary or trend-stationary. The last column of Table I shows the calculated p-values for the unit root tests for each store. Since all values are below 0.05, we can conclude that the data for all stores are (trend-)stationary. Indeed, the only store in Figure 1 for which we observe a structural break is store 12, as previously discussed. Yet, the monthly trend given in orange in that plot is only shifted down the y-axis.

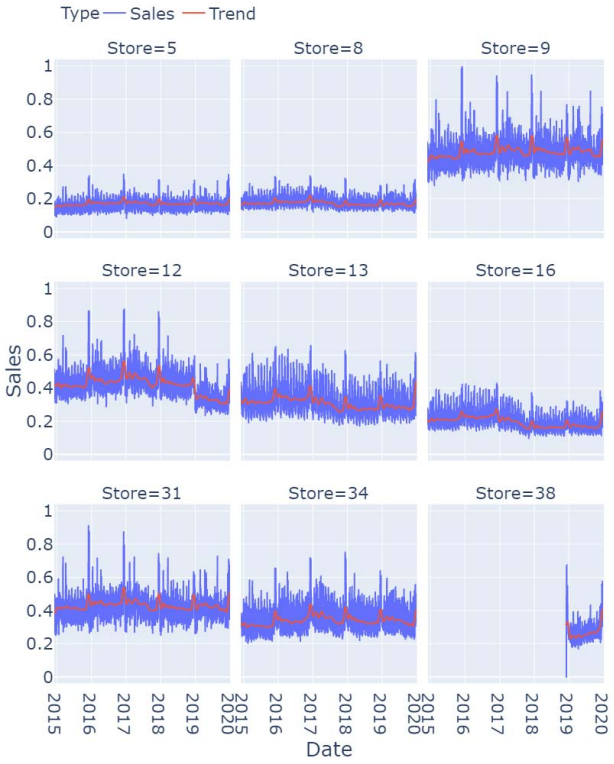

Fig. 1: Per-store total daily sales (blue) and its monthly trend (orange). Values are globally scaled for confidentiality.

3) Inspecting trend and seasonality: We observe seasonality effects for varying periods in the sales data considered. We start our analysis with the monthly trend given in Figure 1 for each store individually, where we see a clearly repeating pattern indicating a strong monthly seasonality. Furthermore, we illustrate in Figure 2 (top) the yearly trend observed for store 9, which we use as representative of the remaining stores, for brevity. The upward increase from 2015-2017 is followed by a plateau between 2017-2019 and a decrease between 2019-2020. This is consistent with macroeconomical indices from the period, as follows. Figure 2 (middle) shows the difference in sales between consecutive years, whereas Figure 2 (bottom) depicts the inflation for the 2014-2020 period. Specifically, each point in Figure 2 depicts the IPCA index accumulated over the 12 months of the given year. The differences between consecutive years for the 2016-2019 range follows the exact shape of the inflation index for the 2015-2018 range.

Besides monthly and yearly, we also observe weekday seasonality, as illustrated with store 9 sales data in Figure 3, comprising selected consecutive weeks from January to February 2019. The hills in the plot represent the weekends, whereas valleys represent working days. This analysis is further detailed for selected stores in Figure 4 using autocorrelation function (ACF) analysis. In the ACF plots, the left-most and the center plots depict daily and monthly autocorrelation,

Fig. 2: Store 9: yearly trend (top) and how the difference between consecutive years compares to inflation index (bottom).

Fig. 3: Store 9: weekday seasonality for selected weeks.

respectively. For brevity, the figure only depicts half of the stores, representative of the remaining as we will later detail. Specifically, from top to bottom Figure 4 respectively gives stores 5, 9, 12, 13, and 16. In the plots, the autocorrelation values vary from -1 to 1, and values that escape the blue shaded area are considered statistically significant. Notice that for the daily ACF plots most values are significant, whereas the opposite is observed for the monthly ACF plots.

Assessing cross-store patterns, we observe that daily autocorrelation follows a cosine-like pattern, though depending on the store group the length of the cosine wave and its range may vary. In detail, for all stores but stores 13 and 16, we observe a high correlation between the given day and its previous corresponding weekdays (lags 7, 14, and so on). To a lesser extent, we also observe correlation with days surrounding the given weekday. Conversely, for stores 13 and 16, the length of the cosine wave indicates a biweekly seasonality. Though not given in Table I, these two stores are located in the same geographical area, for which this biweekly autocorrelation pattern is understandable. Finally, store 12 is the only for which anticorrelation is not observed, which could also be a factor of location, as this is the most central store.

4) Probability distribution: We conclude this exploratory analysis with brief comments on the probability distributions of the sales data from the different stores, given in Figure 4 (right). In general, we observe distributions that are similar to normal distributions, though for some stores we

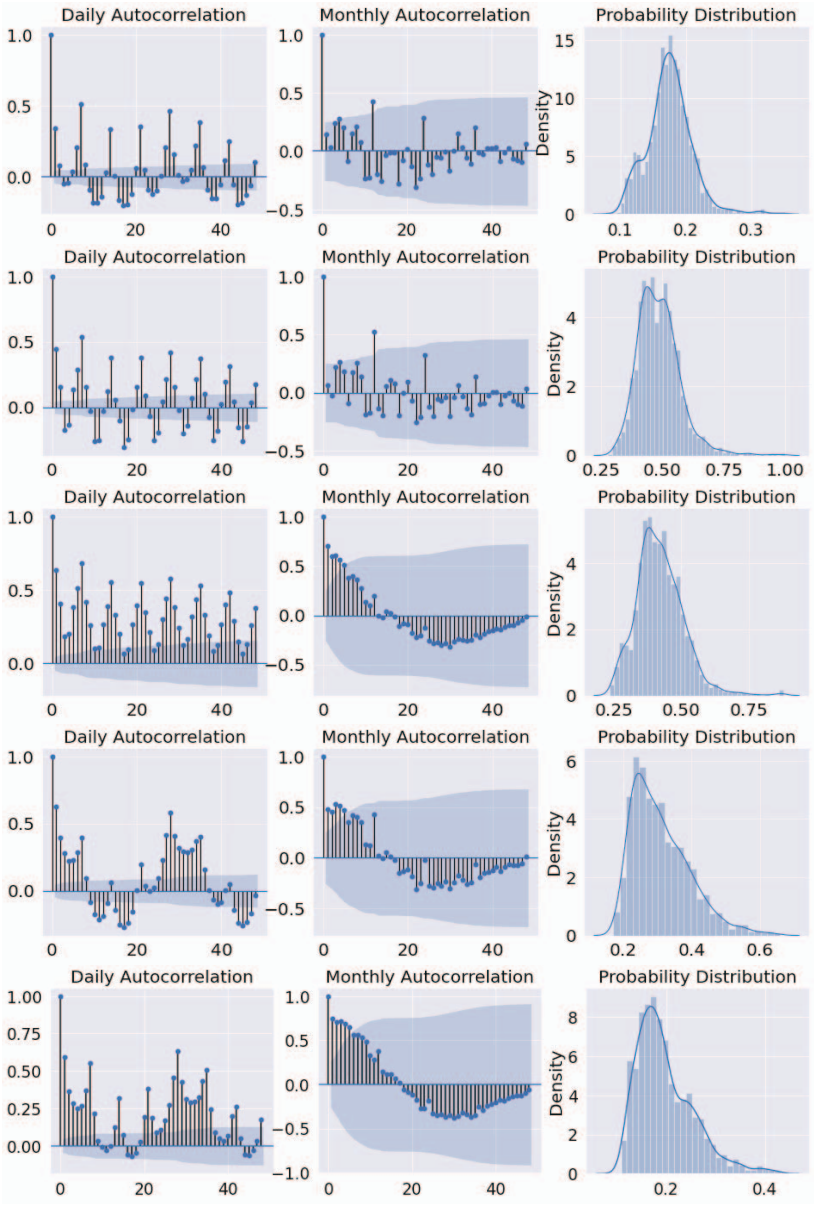

Fig. 4: Daily (left) and monthly (center) ACF and histogram (right) for stores 5, 9, 12, 13, and 16 (from top to bottom).

see varying degrees of positive skewness. Further investigation could also assess in more detail the possibility of bimodality in a few stores, namely stores 5, 9, 12, and 16, though indications for this are very subtle.

In this section, we have focused on the first two steps of the CRISP-DM methodology, namely business and data understanding. Besides grasping important business-related insights, such as the correlation between store size and sales participation, we have observed varying seasonality effects in the data. In particular, the periodicity presented in Figure 2 and further detailed in Figure 4 represents a very important property of the dataset, which will not only guide feature engineering, but also impose an important new baseline estimator. In the next section, we detail data preparation, modeling, and the experimental setup we adopt for evaluation.

IV. Data Preparation and Modeling

The business and data understanding assessment conducted in the previous section produced important insights to guide data preparation and modeling. In this section, we propose our supervised machine learning pipeline to model sales demand from the Nordestão supermarket chain. Initially, we detail

TABLE II: Features engineered for data enrichment.

Type	Features
W&M	weekday_sine, weekday_cosine, month_sine, month_cosine
T&S	monthly_trend, monthly_seasonal, yearly_trend
AR	y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}

the feature engineering we propose, namely the alternative approaches we employ to model seasonality. Later, we define the experimental setup we adopt for modeling and evaluation, namely the machine learning algorithms we select and their training, development, and testing setups.

A. Data preparation and enrichment

Table II lists the features we engineer in this work. Below, we further detail each feature set.

Weekday & month (W&M) are features we engineer in a two-stage approach to account for weekday and monthly seasonality. First, each weekday and month is encoded as ordinal features, respectively ranging from 1 to 7 for weekday and 1 to 12 for month. Then, we produce sine and cosine-transformed versions of these ordinal features to represent the circular relationship between their values.

Trend & seasonal (T&S) components are obtained using the STL method [8]. Specifically, we consider 30-day trend and seasonal components to account for monthly seasonality, as well as 360-day trend for an annual perspective. Since the yearly trend matches the inflation curve observed in the period, we do not enrich the data with macroeconomical indices.

Autoregressive (AR) features are lagged versions of the raw data. A given lagged feature y_{t-k} is the value observed k time periods before a given value y_t. We consider seven lagged features so each point includes data from all previous weekdays, in an alternative attempt to model weekday seasonality.

B. Modeling

As previously discussed, we devise a model for each store, which takes as input a total daily sales series and predicts next-day total sales. As forecasting methods, we select LR as representative of assumption-based algorithms and also the assumption-free RF and XGB algorithms. Our choice is motivated by the effectiveness of these algorithms, as well as the computational infrastructure available at Nordestão. In detail, preliminary experiments with SARIMA models produced results below expectation. In turn, LR is knowingly effective for regression problems when appropriate feature engineering is adopted. Finally, the decision not to adopt deep learning was a business choice based on available resources. To properly evaluate the generalization degree of the models without incurring in data leakage, we adopt a two-stage evaluation approach comprising model development and testing. In detail, at a higher level we use hold-out to isolate the 2020 data, which we use for testing, from the 2015–2019 data, which we use for training and development. We next detail each stage, which commonly consider R^2 as the metric to be optimized.

TABLE III: Different feature subsets used for prediction.

	Set				
Features	**Raw**	**All - W&M**	**All - T&S**	**All - AR**	**All**
Raw data	✓	✓	✓	✓	✓
W&M	—	—	✓	✓	✓
T&S	—	✓	—	✓	✓
AR	—	✓	✓	—	✓

1) Model development: In general, model development comprises (i) feature engineering (FE) assessment and (ii) algorithm configuration. Given the computational infrastructure available at Nordestão, we focus our experimental campaign on the former, adopting default suggested parameters by scikit-learn for RF and by the official implementation of XGB. The experimental design we adopt for FE assessment considers the feature sets given in Table III, and follows a leave-one-out approach. More precisely, the first set comprises only raw data (labeled *Raw*), whereas the last set includes all features engineered (labeled *All*). In turn, the three intermediate sets respectively leave out (i) weekday & month (labeled *All - W&M*), (ii) trend & seasonal (labeled *All - T&S*), and (iii) autoregressive (labeled *All - AR*) features. Effectively, the direct comparison between one these three latter sets and the *All* set indicates the contribution of the left-out feature.

During model development, we employ walk-forward cross-validation on the 2015–2019 data. To keep computational cost constrained, we model folds as years. As such, our assessment considers five folds that are used incrementally over four cross-validation iterations. For instance, at the second iteration the model is trained with the data from 2015–2016 and evaluated with the data from 2017. Besides being computationally reasonable, modeling folds as years rather than as months better reflects the seasonality patterns observed.

2) Model testing: To assess the level of generalization that algorithms achieve when presented with unseen data, we train models on data from 2015–2019, and reserve 2020 data for testing. When testing, we employ two types of analysis:

Single-week. At the lower variability level, a model is evaluated based on a whole-week prediction. Since models are trained to predict next-day sales, we perform whole-week prediction incrementally. More precisely, to predict a given day i of the target week the model has access to all the ground-truth data prior to day i. After total sales for all the seven days of the given week have been individually predicted, we compute the R^2 score between the ground-truth and the predicted week values. With this approach, we reduce the effect of weekday seasonality on results. Note, however, that we do not claim to assess the ability of the models for multi-day prediction. Importantly, we remark that we fit ten RF models to account for its stochastic nature. Hence, when comparing RF with LF or XGB for a given store, RF values are the median R^2 scores of the values predicted by the ten models trained for that store.

Multi-week. At the higher variability level, a model is evaluated based on its whole-week prediction for the 10 initial weeks of 2020. Our rationale is that this setup allows us to

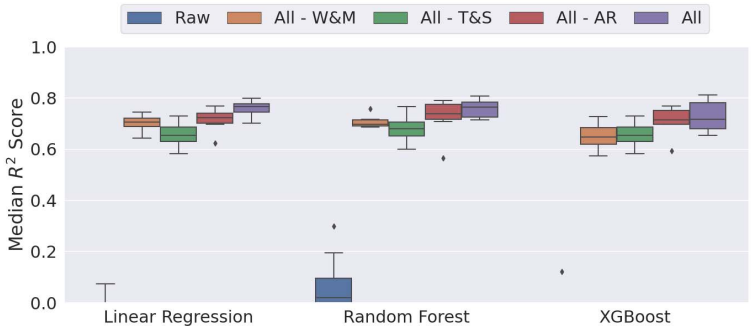

Fig. 5: Median cross-validation performance of the different algorithms when alternative feature sets are provided as input.

observe how model performance varies across the different weeks of the year. This is especially relevant considering that (i) the Carnaval holiday took place in February 2020 and (ii) the first confirmed death by COVID-19 in Brazil occurred during the last week of the period we consider.

Finally, we remark we use two approaches as baseline. The first is the current model applied in the Nordestão chain, namely the average of sales in the previous 45 days (labeled *45DM*, short for 45-day mean). The second baseline comprises the median of the given weekday in the previous 45 days (labeled *45WM*, short for 45-day weekday median). Concretely, this baseline comprises seven estimators that are non-periodic in the scale of weeks. Our rationale is that the weekly seasonality discussed in Figure 4 enforces such a baseline, which should produce a better estimation than the baseline currently in use at Nordestão.

V. Evaluation

The experimental setup detailed above enables an appropriate evaluation of the forecasting pipeline we propose. In this section, we first discuss the contribution of the alternative feature sets engineered. Second, we compare prediction methods from a single-week perspective, also discussing their improvement over baseline approaches. Finally, we discuss multi-week results, highlighting the impact of holidays and social distancing measures on forecasting efficacy.

A. Feature engineering assessment

Figure 5 shows the distribution of the per-store median R^2 scores obtained by estimators during model development using different input feature sets. For clarity, y-axis ranges are clipped to $[0, 1]$, the range of interest for the R^2 metric. All models that use the features we engineer perform better than models that only use raw data, which often obtain negative scores. Interestingly, our feature engineering approach renders even the assumption-based LR forecasting feasible.

When we compare leave-out sets with set *All*, we notice that removing T&S features affects performance much more than removing W&M or AR features. The only exception is XGB, for which W&M proves as important as T&S. As we will discuss later, the similarity in the results distribution between sets *All* and *All - AR* indicates that autoregressive features could be discarded in favor of model simplicity. However, the remainder discussion we conduct in this section considers models that used all the features we engineered.

B. Single-week analysis

Figure 6 gives the performance of each algorithm for stores altogether (top left plot) and individually (remaining plots) when using all features to predict the week comprising January 9th to 15th, 2020. We choose this week for our initial analysis as it is the first week in 2020 for which autoregressive features would not include the January 1st holiday, when all store chains were closed and no sales were made.

The overall performance analysis shows that the forecasts using ML models outperform the baseline in use at the company (45DM). Indeed, the boxplot of the R^2 scores obtained by this baseline does not appear in the plot, since its values are negative. This is an effect of the period of the year depicted, as December holiday season shopping produces peaks and valleys that averages cannot predict well. On the other hand, the alternative baseline (45WM) is much more competitive, addressing to some extent the weekday seasonality effects previously discussed. Yet, the high variance observed in the results indicates that the performance of 45WM is strongly affected by store characteristics.

Regarding ML algorithms, we see that the first and third quartiles of RF and XGB are very similar, with both algorithms sometimes achieving R^2 scores above 0.8, and never below 0.4. By contrast, the distribution for LR is lower on the y-axis, and a strong outlier is clipped from the plot. Yet, when we assess median performance, we see that RF performs best, followed by LR, and lastly XGB. When both ML algorithms and baselines are jointly considered, we see that median performance of 45WM is only better than the median performance of 45DM (though by a significant margin).

We then proceed to the store-by-store analysis (remaining plots in Figure 6), where we remark that RF results are given as medians of the ten repetitions conducted, as previously discussed. For each store considered, at least one of the ML methods achieves a (median) score of 0.5 or above, often higher than 0.8. As previously discussed, the baseline in use

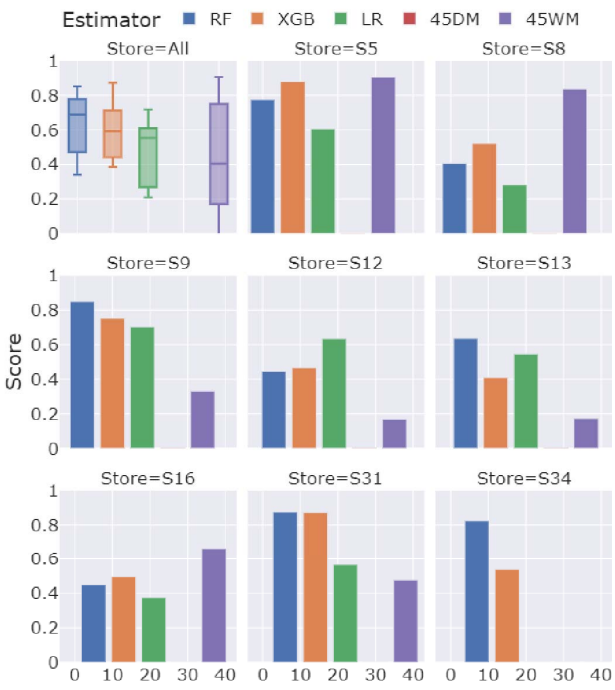

Fig. 6: Overall (top left) and per-store (remaining) single-week prediction performance of ML algorithms, compared to different baseline approaches.

at the chain (45DM) achieves negative scores. On the other hand, the alternative baseline (45WM) can achieve scores closer to ML models, even outperforming them for the smallest stores (5, 8, and 16). This result is in line with the plots given in Figure 1, which show that that these stores do not present a large variation in sales on special dates, e.g. holidays or seasonal periods, in order to bring about relevant sales peaks. The worst performance displayed by 45WM concerns store 34, that comprises a strong outlier for multiple estimators, and which we intend to further investigate in future work.

Comparing ML algorithms, each of the ensembles outperformed the remaining algorithms for half of the stores considered. In turn, LR presented contrasting results, sometimes better than one or either ensembles (stores 12 and 13), but often worse than the ensembles and/or WM45. Regarding stores 12 and 13, none of the ML algorithms produced scores higher than 0.65. Still, we consider these results reasonable, especially given the previously discussed structural break incurred on store 12 data by the recently inaugurated store 38.

C. Multi-week analysis

Figure 7 gives results for a 10-week window of prediction for all stores starting on the first working day of 2020 and ending just before COVID-19 quarantine. We remark that (i) the models were not retrained during testing, and; (ii) results are grouped by week (top) or store (bottom). On the top plot, the day given under a group of boxplots indicates the initial day of the week considered. Note that the boxplots for the week starting on January 9th were already previously discussed, and serve as baseline for our following discussion.

When we assess the variations in results along the weeks, we see that forecasts get worse in the weeks after or surrounding a holiday. In detail, the week starting in January 2nd is affected by New Year's eve, whereas the weeks starting in February 13th, 20th, and 27th surround Carnaval. In addition, the week starting on March 5th reflects the increasing fear incurred by the pandemic, made stronger by the first confirmed death by COVID-19 in Brazil. For the other weeks, forecasts are reasonably effective, although variance can be observed either as a factor of stores or prediction algorithm. Importantly, 45WM remains competitive with ML algorithms, whereas boxplots for 45DM remain mostly clipped out of the plot.

When we assess results grouped by store, we see that the most often median scores obtained by the ML algorithms remained between 0.6 and 0.8. A few exceptions concern stores 8, 9, and 12, which had greater variations and/or lower scores. This is in part justified by the variations in weeks, previously discussed. Nonetheless, it is remarkable that the results from the remaining stores be so high, especially for the ensemble ML algorithms. This is especially relevant when we see that for some of these stores 45WM presented high variance and/or low scores.

VI. Further Analysis

The initial evaluation discussed in the previous section demonstrated the efficacy of the features engineered and the impact of store and holidays on model performance. To conclude our assessment, we briefly (i) evaluate whether a single ML algorithm could be recommended for Nordestão; (ii) investigate the benefits of exogeneous features to model holidays, and; (iii) demonstrate how models could be further simplified. Importantly, this additional analysis not only deepens our understanding of the results and points future work directions, but is also instrumental for model deployment and maintenance, the final step of the CRISP-DM methodology.

Algorithm recommendation. An overall comparison of the ML algorithms and baseline approaches considered in this work across stores and weeks is given as a boxplot in Figure 8. In a nutshell, these results confirm what has been previously discussed, namely (i) the similarity between the distributions of RF and XGB; (ii) that the distribution of LR is somewhat between the distribution of ensemble algorithms and 45WM, and; (iii) the very poor performance of 45DM. Altogether, the intersection in distribution between most algorithms render differences non-significant according to Friedman's chi-squared test. Yet, RF (171) and XGB (180) are the best-ranked algorithms, followed by LR (230) and 45WM (250). In this context, we recommend that RF and XGB models be deployed interchangeably at Nordestão, as a function of their previously discussed store-wise performance.

Exogenous features. Given the effect of holidays on results, we investigate whether further enriching the dataset would

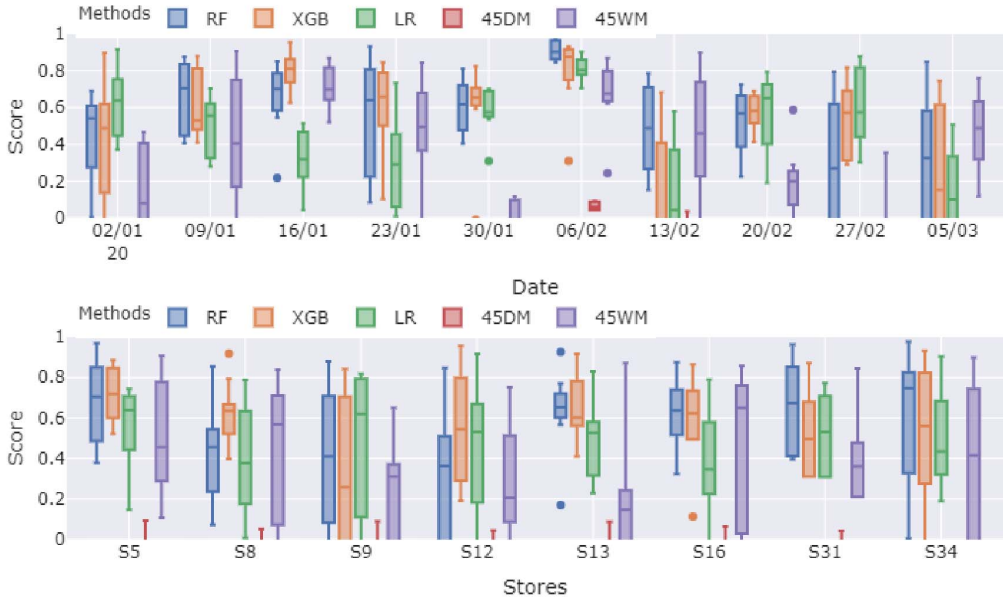

Fig. 7: Multi-week performance of the different algorithms for all stores, when grouped by week (top) or by store (bottom).

Fig. 8: Overall algorithm comparison across weeks and stores.

improve model performance. Our rationale is that some holidays are very important, but their exact dates may vary over the years. Concretely, we model holidays as a binary feature, where a non-null value indicates that the given day was a holiday in the given year. Figure 9 shows the comparison between model performance under the development setup when taking alternative input feature sets. For this analysis, we limit our discussion to *All* and *All + Holidays*, since the latter combines all endogenous features from the former with the manually-enriched holiday binary feature. In general, the R^2 score distributions from these two feature sets are very similar, but median values using *All + Holidays* are improved for all algorithms. For brevity and model simplicity, we list a deeper analysis on holiday features as future work. Yet, we remark that holiday modeling could likely benefit from an encoding that allow models to relate different dates in different years as the same holiday (e.g. Carnaval and Easter).

Model simplicity. As previously discussed, some of the benefits provided by the features engineered in this work could be traded for model simplicity, if so desired by Nordestão. The first set comprises the autoregressive features, as mentioned in Section V. The second is the binary holiday features, as detailed above. Here, we further investigate a third and last feature set that could be likely removed for model simplicity. Specifically, we ablate the *W&M* feature set to understand if both weekly and monthly features contribute to model performance. Our rationale is that Figure 4 indicated a strong weekday autocorrelation, but monthly not as much. This analysis is given in Figure 9, which depicts the performance under the model development setup of sets *All - Weekly* and *All - Monthly*, which are obtained by respectively leaving out weekly and monthly features. Overall, discarding the monthly features has little to no effect on model performance, whereas discarding weekly features worsens results considerably.

VII. Conclusion

Retail sales forecasting [6], [14] is a topic widely explored in the industry for its role in strategic, tactical, and operational decision-making [1]. However, the existing literature mostly comprises case studies [10], [11], [14], [17], [19], and proper algorithm selection and data preparation requires a deep understanding of the business and its data. In this work, we addressed sales forecasting for the Nordestão supermarket chain, one of the largest in Brazil [7]. Specifically, we have produced forecasting models using effective machine learning methods for eight of their retail stores. To improve the performance of these models, we have engineered features using alternative approaches to handle varying seasonal effects observed. Our assessment demonstrated the (i) benefits of the features engineered; (ii) the importance of considering multiple prediction algorithms and evaluating models over time and across stores, and; (iii) the improvements over the current forecasting approach employed at Nordestão.

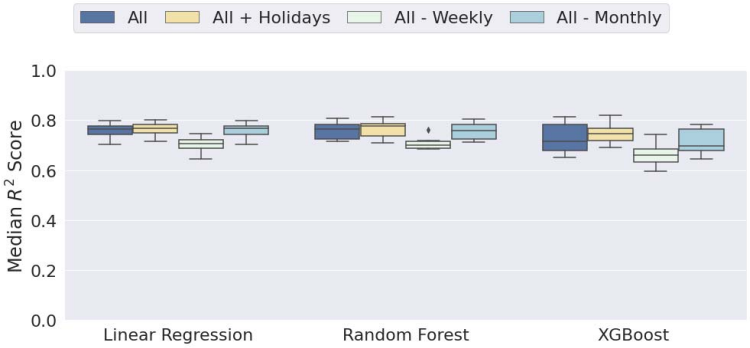

Fig. 9: Median cross-validation performance of the different algorithms when alternative feature sets are provided as input.

Though our paper addresses the particular scenario of the Nordestão supermarket chain, the main findings we observe when developing a sales forecasting model for retail can be reused by other companies. As expected, data exploration confirmed important seasonal effects, namely monthly and weekday. We also noted that the COVID-19 pandemic has completely altered the behavior of sales, which require an incremental modeling approach starting from pre-pandemic data. Another important aspect of our work is how feature engineering transformed the nonlinear data in a way that was compatible with linear models such as linear regression (LR). Furthermore, adopting a variety of models enables portfolios, delivering even more robust results. Finally, we confirmed how baseline models currently employed by the industry can be overly naive, thus significantly impairing planning. Besides showcasing an improved baseline, we have demonstrated how assumption-free models are a better choice for the industry.

Our investigation opens a number of important future work possibilities. First and foremost, the tactical-level forecasting we propose is seminal for operational-level approaches that can improve stock planning and distribution. A second, challenging possibility is to (i) assess and (ii) mitigate the impact of the COVID-19 pandemic on sales. Regarding (i), this will likely require not only daily, but also monthly forecasting. Concerning (ii), a direct consequence of the pandemic is the rupture in the series from all stores, compromising models targeting the current period. Addressing this issue is paramount to improve the current forecasting ability of the chain. Finally, our literature review indicated the lack of a meta-methodology to be used in forecasting for the retail sector, empirically verified. In this context, we plan to expand our work to point out from a general perspective the main aspects to address in every step of a forecasting retail sales pipeline, as well as the main challenges and alternatives to tackle them.

REFERENCES

[1] R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," *Int. J. of Forecast.*, 2019.

[2] J.-H. Böse, V. Flunkert, J. Gasthaus, T. Januschowski, D. Lange, D. Salinas, S. Schelter, M. Seeger, and Y. Wang, "Probabilistic demand forecasting at scale," *VLDB*, vol. 10, no. 12, pp. 1694–1705, 2017.

[3] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *KDD*. IEEE, 2016, pp. 785–794.

[6] A. L. Loureiro, V. L. Miguéis, and L. F. da Silva, "Exploring the use of deep neural networks for sales forecasting in fashion retail," *Decisi. Support Syst.*, vol. 114, pp. 81–93, 2018.

[7] Associação Brasileira de Supermercados, "Ranking ABRAS," https://www.abras.com.br/edicoes-anteriores/Main.php?MagNo=259, 2020, accessed: 2021-05-06.

[8] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "Stl: A seasonal-trend decomposition," *J. of Official Stat.*, vol. 6, no. 1, pp. 3–73, 1990.

[9] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Springer-Verlag London, UK, 2000.

[10] C.-Y. Chen, W.-I. Lee, H.-M. Kuo, C.-W. Chen, and K.-H. Chen, "The study of a forecasting sales model for fresh food," *Expert Syst. with Appl.*, vol. 37, no. 12, pp. 7696–7702, 2010.

[11] Q. Wen, W. Mu, L. Sun, S. Hua, and Z. Zhou, "Daily sales forecasting for grapes by support vector machine," in *CCTA*. Springer, 2013, pp. 351–360.

[12] L. R. Berry, P. Helman, and M. West, "Probabilistic forecasting of heterogeneous consumer transaction–sales time series," *Int. J. of Forecast.*, vol. 36, no. 2, pp. 552–569, 2020.

[13] L. Aburto and R. Weber, "Improved supply chain management based on hybrid demand forecasts," *Appl. Soft Comput.*, vol. 7, no. 1, pp. 136–144, 2007.

[14] X. Liu and R. Ichise, "Food sales prediction with meteorological data—a case study of a japanese chain supermarket," in *DMBD*. Springer, 2017, pp. 93–104.

[15] Z. Wang, T. Hong, H. Li, and M. A. Piette, "Predicting city-scale daily electricity consumption using data-driven models," *Advances in Applied Energy*, vol. 2, p. 100025, 2021.

[16] G. Tsoumakas, "A survey of machine learning techniques for food sales prediction," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 441–447, 2019.

[17] F. C. de Almeida and A. F. L. Passari, "Previsão de vendas no varejo por meio de redes neurais," *Revista de Administração*, vol. 41, no. 3, pp. 257–272, 2006.

[18] C. F. de Angelo, R. Zwicker, N. M. M. D. Fouto, and M. R. Luppe, "Séries temporais e redes neurais: uma análise comparativa de técnicas na previsão de vendas do varejo brasileiro," *Brazilian Bus. Rev.*, vol. 8, no. 2, pp. 1–21, 2011.

[19] S. Brusque and L. C. Zucatto, "Previsão de vendas para empresa varejista de confecções adulto feminino e masculino," *Revista de Administração e Negócios da Amazônia*, vol. 7, no. 2, pp. 88–111, 2015.