International Journal of Combinatorial Optimization Problems and Informatics, 13(2), May-Aug 2022, 33–39. ISSN: 2007-1558

www.editada.org

Data-Driven Enterprises for Economic Recovery after Pandemic. A Study Case in Consumer-Packaged Goods Industry

Roberto Contreras-Masse[1,2*], *Alberto Ochoa-Zezzatti*[1], *Luis Pérez-Domínguez*[1], *Karen Naudascher*[3]
[1] Universidad Autónoma de Ciudad Juárez, Cd. Juárez, Chihuahua, Mexico
[2] Tecnológico Nacional de México, Cd. Juárez, Chihuahua, Mexico
[3]University of Florida, Florida, USA
*Correspondance: rcontreras@itcj.edu.mx

Abstract. Businesses across Latin America are working towards economic recovery. This case study of the Consumer-Packaged Goods (CPG) industry shows how business intelligence can help organizations more accurately forecast demand by applying business analytics and machine learning. The study case discusses how traditional forecasting methods are being utilized today and what limitations and challenges they are facing, and what machine learning alternatives exist. Also, it is discussed why data granularity was a key factor in creating the proposed forecasting models and how machine learning capabilities provide better insights and more accurate results. Lastly, the implementation of data-driven philosophy is discussed, explaining components, interaction, data gathering, transformation, and what technologies were used. This comprehensive study case can help other enterprises looking to start their journey to become data-driven organizations relying on business analytics to make better decisions during their efforts for economic recovery.

Keywords: Business Analytics, Decision Making, Machine Learning, Business Intelligence, Data Pipelines.

Article Info

Received: September 1, 2021
Accepted: November 29, 2021

1 Introduction

The pandemic developed new behaviors from work and learning to live at home and shopping patterns. The industry of Consumer-Packaged Goods (CPG) is a multibillion-dollar market that was impacted by the COVID-19 pandemic during 2020. In the United States, overall consumption declined 15%; India reported a decrease near 7.5% [1]; Romania reported 46% of consumers with financial struggles [2]. Retailers are the principal channel for CPG products, and those retailers had to reevaluate how their stores operate, being affected by consumer habits changes [3].

To have a faster economic recovery, CPGs are looking to different business models and different markets. Consumers are increasingly attached to smartphones as their main computer interface for shopping [4],[5]. As consumer habits change, CPGs must change strategies to survive. The main strategy is the optimization of processes to reduce expenses and maintain operations [6]. Also, CPGs have expanded to different products and market segments. As an example, soda bottlers produce and distribute snacks and oat goods; brewers have begun to offer water products marketed as seltzers; other companies have stores open to the public for direct sales.

Direct sales businesses require demand planning and sales forecasting, among other budgetary activities. If the physical goods produced have a wide range of options and presentations, it becomes complex to convey these budgetary activities in an effective manner. It is common for these companies to rely on spreadsheets to crunch their numbers [7]. However, data gathering, data transformation, analytics, and decision-making consume time and prevent the results from being used in a timely manner.

© Editorial Académica Dragón Azteca S. de R.L. de C.V. (EDITADA.ORG), All rights reserved.

This paper is organized in sections. The Background Section presents the study case used in this research. Then, the methodology used in this research is presented. The next section discusses the results and findings. Finally, the next steps and future research are exposed.

2 Background

Forecasting techniques have been studied during the last century with increasing reliability. One of the techniques in forecasting commonly used in retail sales industries is time series forecasting [8],[9]. A time series contains data points indexed in time order. Usually, an index uses a timestamp that has a date-part and a time-part, and, on some occasions, it can include the time zone and milliseconds. The index can be later grouped by coarse-grain dimensions, such as minutes, hours, days, up to months, quarters, and years [10]. Product sales can also be modeled as a time series problem since it is common to have the appropriate level of granularity (time of sale) from the point-of-sale system. In addition, time-series are useful to predict values within a period for an entity i at a time t, taking the form of $y_{i,t}$ and the simplest forecast models are represented by [10]:

$$\gamma_{i,t+1} = \rho(\gamma_{i,t-k:t}, \theta_{i,t-k:t}, \sigma_i) \tag{1}$$

γ are the forecasted values based on θ observations looking back k periods in time t and taking into consideration statistical metadata σ.

There are different techniques and models for time series forecasting that have emerged from academic research, including, but not limited to, exponential smoothing, autoregressive and moving average, seasonality, state-space models, nonlinear models, long memory models, autoregressive conditional heteroscedastic models, count data, and Gaussian processes regression [11],[12]. Having forecasting as the first challenge faced by CPGs, accuracy is the other challenge, and it is often a metric well accepted to indicate if the forecasting model was effective or should be improved. Accuracy is usually measured by statistical error indicators with several variants, as discussed by Gooijer and Hyndman [11].

Selecting the appropriate model depends on data behavior; thus, CPG products tend to behave in seasonal patterns, with organizations choosing models that focus on sales or demand as the primary metric. In other studies, specific to demand or sales forecasting for retail, the studied forecasting models rely on Autoregressive Moving Average (ARIMA) and its variants [13]. In recent research, adding machine learning (ML) for forecasting has been on the rise. The use of ML algorithms such as generalized linear models, decision trees, and gradient boost trees have produced reliable data-driven results in retail [14].

Most of the studies use one single version of a predictive model, while others add variables to a single focus area. As an example, the study case for electricity demand has been leveraged many times, and other authors have introduced new variables such as wind, temperature, or season, trying to answer the same question: what is the electricity demand? CPGs can have multiple variants of the same item to forecast; therefore, it creates a complex problem to forecast each of the variants. For those cases, one approach accepted is to treat each variant as a single item and forecast each case individually [15]. In CPG products, there could be a high number of variants, shifting the problem to analysis of very large amounts of data, or nowadays known as big data.

2.1 Case Study

One of the largest CPG companies in Latin America is a leader in products sold directly at their almost 300 stores across the country, offering 53 variants of the same product. In 2020 sales were impacted due to the pandemic, with the sales volume decreasing from 2019 by 28%. As part of their strategy for economic recovery, the sales forecast is required to optimize distribution and production and to evaluate new business models to help revenue and margin increase. For academic purposes, let's call this company CPG1, the product to work on P and their variants can be called flavors F_i, where $i = \{1..53\}$; the stores will be referenced as S_j, where $j = \{1..278\}$.

CPG1 currently has a good understanding of the sales in coarse granularity, meaning they analyze sales by location and product line, but not by variations (or flavors). They use spreadsheets to calculate the demand for the upcoming week and plan the production and distribution accordingly using moving average models. Data gathering and transformation is a challenge, as information comes from different sources in different formats. Delivery of results is time-consuming as there is a high amount of manual labor. The methodology proposed in this research aims to address these issues and optimize the decisions for the supply

chain based on daily sell-out. The goal is to gain efficiencies in supply chain demand, achieving transformation into a data-driven organization with automated delivery of insights.

3 Methodology

The methodology proposed in this research consists of five steps as follows. First, gather available historical data and identify key data points to be included in a forecasting engine every day. Daily feed frequency is limited by point of sale (POS) capacity and cannot be improved unless changing their POS vendor, which would be counterproductive to the primary goal of improving margins. The second step is data transformation to be in a time-series format. Among transformations, it is required to extract only POS transactions regarding product P. Also, if there are promotions or free samples given away, those marked by POS reports should be discarded.

The third step consists of time series analysis to understand trend, seasonality, and auto-correlation to prepare for a forecasting algorithm. The fourth step is to select the best ARIMA and machine learning models. Finally, train the algorithms and test against test data sets to determine accuracy. The results provide the best option to be deployed for daily forecasting.

4 Results

The historical data available corresponds to the years 2019 and 2020; however, the year 2020 was an atypical year due to the pandemic. CPG1 was forced to temporarily close some stores producing a significant loss of sales. Therefore, the most consistent data for sales forecasting for the purpose of this study is 2019. The convenient place to store this data for analysis is in the cloud for easy access. This space serves as a data lake and stores the daily data feeds using data pipelines.

4.1 Analysis of 2019 data

Decomposition analysis in Figure 1 shows seasonality with increasing and decreasing trends. Because there is seasonality, Seasonal Auto-Regressive Integrated Moving Average with eXogenous (SARIMAX) is suggested for forecasting.

Fig. 1. Decomposition results for time series 2019

In the auto-correlation graph, it suggests lag=7. This is interesting as product P is purchased more during weekends and Wednesdays due to special offers. However, the analysis by wholesales per day is not providing the real behavior of a particular store. Therefore, because the data source is so granular, it is possible to execute forecasting on one store for one flavor and understand the behavior in a personalized fashion.

Data needs to be transformed to allow better handling. All flavors were standardized to avoid multiple labels for the same flavor, e.g., apple and Apple are the same flavors. Then, all flavors were masked to avoid any bias when selecting the sample sets. A timestamp field was created from the data, and this field was promoted to "Date Time" index. Remember, date-time indexes can be searched and presented in assorted dimensions, such as yearly, quarterly, monthly, or by hour, or by minute, among many other options.

4.2 Analysis of Single Flavor for One Location

The initial analysis selects a random store. Let, $S = \{156\}$, noted as S_{156} and a random flavor F_{15} to run SARIMAX vs. XGBoost forecast. The pattern of one flavor sales during the year is shown in Figure 2, where the x-axis shows the day of the year, and the y-axis is the summary of sales. There are 320 data points found within the data set, with $\mu = 6.12$, $\sigma = 4.14$, the minimum sale is one item, highest sales are 22 items. Taking another random store, it is obvious there are different sales quantities. For instance, store S_{14} sold 176 items of the same flavor, with $\mu = 9.32$, $\sigma = 7.44$, lowest sale of one item, and top sale of 37 items.

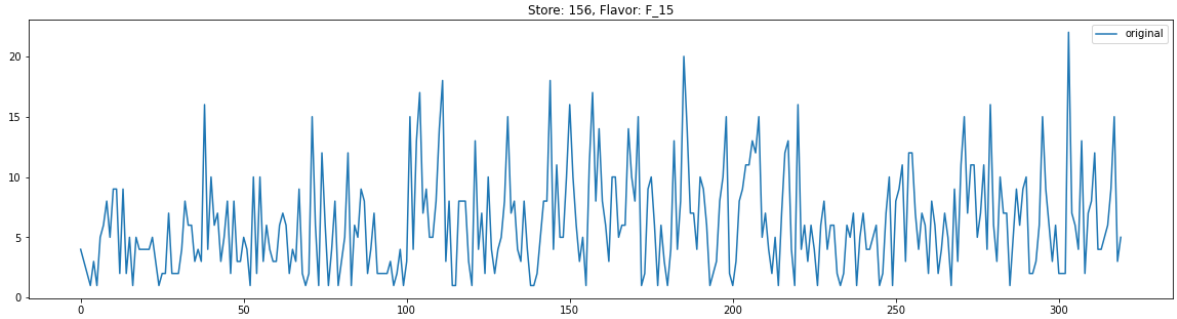

Fig. 2. Behavior of 2019 sales for flavor F_{15} in store S_{156}

It can be observed each store has a specific sales pattern that varies among flavors. This sales pattern does not allow for accurate forecasting of demand if data is aggregated by net sales. If the business is looking for sales volume, no matter what flavors are sold, then aggregation could be handled in simple tools such as spreadsheets. However, it is almost impossible to forecast a more precise demand accounting for variabilities such as flavor, location, season, etc. The market is different for each location, as can be seen in Table 1 that shows a sample of sales of flavor F_{15}.

Table 1. Sales at random store sample $n = 10$ of flavor F_{15}

Store	Sales
75	3,345
138	2,235
157	2,175
43	1,267
201	1,521
69	617
26	1,945
189	587
77	3,371
56	783

The best approach is to forecast flavor per location; this approach enables the accurate demand of a mix of flavors needed per location. Also, as each location belongs to a city, region, or route, it is feasible to forecast the complete demand of a city, region, or route, optimizing production and distribution.

4.3 Forecasting Results

Forecasting was evaluated with two methods. The results for SARIMAX were encouraging. SARIMAX optimizer found the best model as ARIMA (1,0,0) x (1,0,0,7) in 15.016 seconds. SARIMAX training set and test set use three quarters and one quarter, respectively. The accuracy of the SARIMAX model was 85.12%, which is acceptable.

The same subset forecasted with XGBoost Regressor provided an accuracy of 96.04%. Figure 3 shows the real sales recorded on store 156 for flavor 15 (blue line), and XGBoost estimation is represented by the orange line. It is clear the XGBoost estimation is very close to the real numbers.

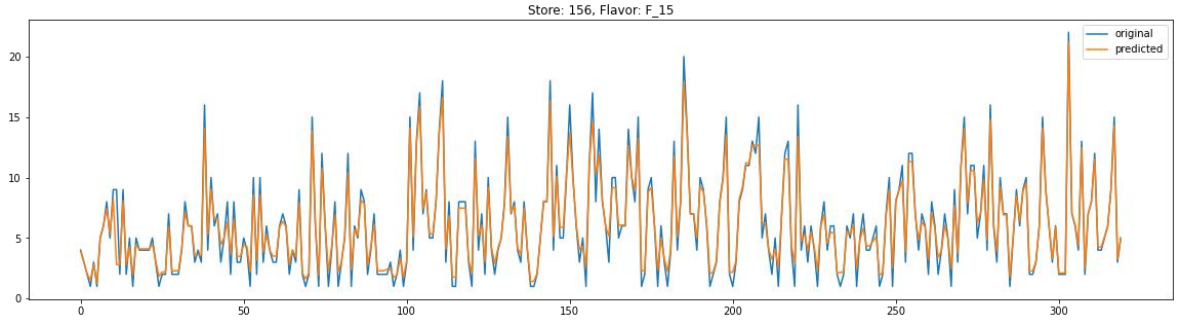

Fig. 3. XGBoost forecast vs. real for flavor F_{15} in store S_{156}

The same stores selected in Table 2 were forecasted with both methods, and the best-performing forecast method is identified in the last column.

Table 2. Forecast performance comparison at random store sample $n = 10$ of flavor F_{15}

Store	SARIMAX	XGBoost	Best (S or X)
75	0.85	0.96	X
138	0.78	0.99	X
157	0.81	0.98	X
43	0.77	0.96	X
201	0.82	0.99	X
69	0.84	0.98	X
26	0.79	0.98	X
189	0.82	0.94	X
77	0.81	0.95	X
56	0.78	0.98	X

XGBoost Regressor surpasses SARIMAX in all cases. The sample provides a good insight to use XGBoost as the selected method for forecasting. The total number of models $[S\ X\ F]$ evaluated was 14,734. The next step was to compare all models with both methods and obtain the average accuracy as a simple metric. The results were SARIMAX = 0.8134, while XGBoost was 0.9657. As a side note, the SARIMAX model fit for all cases took 0.19s; XGBoost used 0.65s. However, finding the best SARIMAX parameters consumed 15.42s. That means, when the model will need to be retrained, it will potentially require parameter tuning. Figure 4 compares the real or true values (gray line) against the estimated values by XGBoost (orange dots) and the estimations using SARIMAX (blue x's).

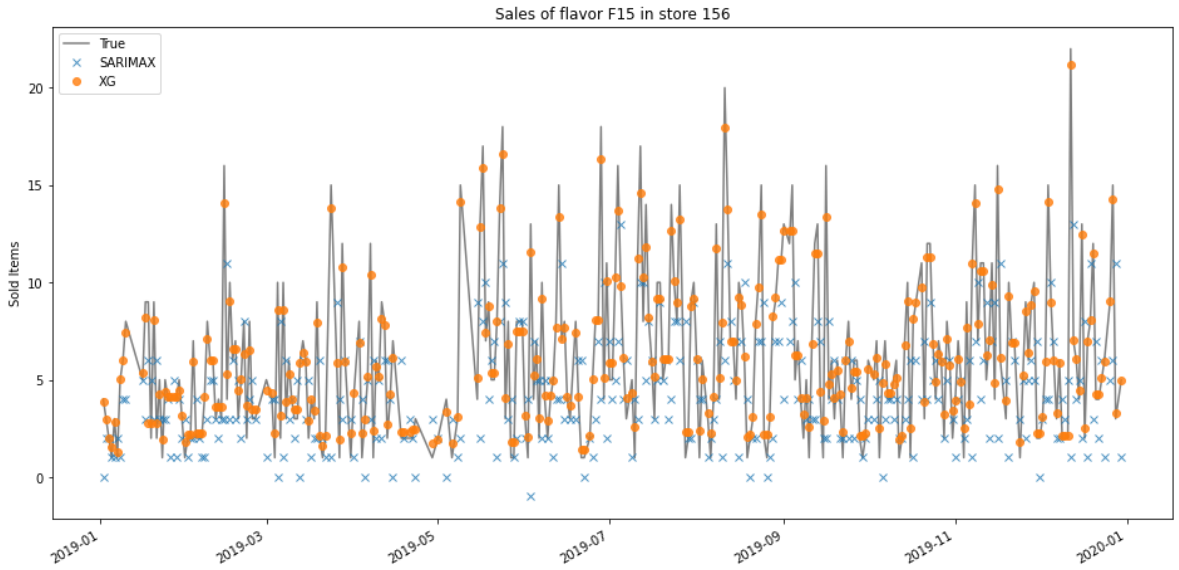

Fig. 4. XGBoost and SARIMAX forecast comparison vs. true values for flavor F_{15} in store S_{156}

4.4 Improve XGBoost Accuracy

As XGBoost needs a seed as a starting point, it is expected to receive different results for different seeds. To find the best-trained model for each case $[F_i, S_j]$ it was proposed to run ten times XGBoost fit for every case with randomized seeds to find the highest accuracy model. Models with 100% accuracy were discarded. To speed up the process, a target accuracy was defined, thus as soon as the accuracy reached the target, the process stopped and did not look for another solution as $Target \leq Acc < 1$.

The result of this process with a target accuracy of 98% was reached in 13,876 of 14,734 cases, corresponding to 94.79%. The 858 times where target accuracy was not reached showed the lowest accuracy of 55.55%; there were two cases that can be considered an outlier with a zero accuracy. Further observations unveiled those stores were closed or moved. 25th percentile marked an accuracy of 95.38%, 50th percentile was 98.02%, and 75th percentile marked 99.45%. The highest accuracy mark was 99.89%.

Table 3. Performance percentiles of accuracy observed on XGBoost forecasting

Percentile	Accuracy Observed
25	95.38%
50	98.02%
75	99.45%
Max	99.89%

5 Conclusion and Future Research

The results of this research suggest a much better accuracy performance from XGBoost over SARIMAX. Approaching the demand problem as individual cases provides more comprehensive insights when compared against solely relying on aggregated sales. It is feasible to obtain a good description of product and variation demand by combining computing power with the appropriate granularity.

Even though SARIMAX has been used in the past with good results, as reported by literature, this experiment, along with the results, let researchers explore alternative machine learning techniques to solve demand forecast problems, more accurate but with much more data needed, such as XGBoost algorithm. It has been observed traditional spreadsheets have capacity limitations preventing proper handling of all granular data. Even so, spreadsheets are and can be extremely useful for professionals not familiar with python or R Language. Despite the present popularity of businesses using spreadsheets to forecast demand, it is clear

CPGs could benefit from machine learning to improve their data analytics and data insights allowing more precise demand forecasting.

As data flows daily to the forecasting system, it is possible to continuously train the algorithm to calculate the most accurate predictions. A daily data feed is more than enough to have information available for updated forecasting. It is clear the update frequency (data ingestion) and business questions will drive the frequency of model updates. For demand forecasting in CPG, product inventories once a day correctly fit the business needs.

The individual product flavor at individual store forecasting provided a production improvement and distribution savings of nearly 3.2% cost in logistics. In addition to this optimization, the product expiration and leftovers were reduced by 12.6%, translating to direct savings for productions. By knowing what stores sell what products, the assortment of flavors was also optimized, although it was not reported for the research. Different industries would need to ingest data at varying paces. Manufacturing, where one of the common problems to solve for is the supply-demand and equipment wearing of energy consumption, could require information by the second or microsecond.

Further research is still open on how these forecasting tools can help to increase sales by personalizing flavors. This marketing impact is an opportunity area to expand the usage of machine learning in forecasting and other optimization challenges. In addition to exploring different business areas to apply these techniques, there are many optimization challenges that could be solved or improved within the CPG industry. By having the demand forecasted and grouped by regions or routes, businesses could optimize the supply chain for product distribution, identifying the best route with the best truckload to reduce the expense in fuel consumption, for instance, allowing the right product mix in the right location delivered in the right time.

References

1. Y. Mahajan, "Impact of Coronavirus Pandemic on Fast Moving Consumer Goods (FMCG) sector in India", *Journal of Xian University of Architecture & Technology*, vol. XII, no. IX, 2020; http://xajzkjdx.cn/gallery/5-sep2020.pdf
2. S. Stanciu, R. I. Radu, V. Sapira, B. D. Bratoveanu, and A. M. Florea, "Consumer Behavior in Crisis Situations. Research on the Effects of COVID-19 in Romania", *Annals of the University Dunarea de Jos of Galati*: *Fascicle: I, Economics & Applied Informatics*, vol. 26, 2020; https://doi.org/10.35219/eai1584040975
3. S. Kohli, B. Timelin, V. Fabius, and S. M. Veranen, How COVID-19 is changing consumer behavior–now and forever, McKinsey & Company, 2020; https://v.fastcdn.co/u/c81ab06a/53497572-0-how-covid-19-is-chan.pdf
4. F.-V. Pantelimon, T.-M. Georgescu, and B.-Ş. Posedaru, "The impact of mobile e-commerce on gdp: A comparative analysis between Romania and Germany and how covid-19 influences the e-commerce activity worldwide", *Informatica Economica*, vol. 24, p. 27–41, 2020; https://doi.org/10.24818/issn14531305/24.2.2020.03
5. M. Schrage, "Data, Not Digitalization, Transforms the Post-Pandemic Supply Chain", MIT Sloan Management Review, 2020; https://sloanreview.mit.edu/article/data-not-digitalization-transforms-the-post-pandemic-supply-chain/
6. S. Li, "How Does COVID-19 Speed the Digital Transformation of Business Processes and Customer Experiences?", Review of Business, vol. 41, p. 1–14, 2021; https://usiena-air.unisi.it/retrieve/handle/11365/1127888/340590/Review-of-Business-41%281%29-Jan-2021.pdf#page=5
7. S. L. Smith, A. S. Golden, V. Ramenzoni, D. R. Zemeckis, and O. P. Jensen, "Adaptation and resilience of commercial fishers in the Northeast United States during the early stages of the COVID-19 pandemic", *PloS one*, vol. 15, p. e0243886, 2020; https://doi.org/10.1371/journal.pone.0243886
8. Al Mamun, M. Sohel, N. Mohammad, M. S. H. Sunny, D. R. Dipta, and E. Hossain, "A comprehensive review of the load forecasting techniques using single and hybrid predictive models", *IEEE Access*, vol. 8, p. 134911–134939, 2020; https://doi.org/10.1109/ACCESS.2020.3010702
9. G. Rivera, R. Florencia-Juárez, J. P. Sánchez-Solís, V. García, and C. D. Luna, "Forecasting the Demand of Parts in an Assembly Plant Warehouse Using Time-Series Models", *POLIBITS*, 2020, vol. 62, p. 59-67; https://doi.org/10.17562/PB-62-7
10. B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey", *Philosophical Transactions of the Royal Society A*, vol. 379, p. 20200209, 2021; https://doi.org/10.1098/rsta.2020.0209
11. J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting", *International Journal of forecasting*, vol. 22, p. 443–473, 2006; https://doi.org/10.1016/j.ijforecast.2006.01.001
12. G. Jeong, S. Park and G. Hwang, "Time Series Forecasting Based Day-Ahead Energy Trading in Microgrids: Mathematical Analysis and Simulation", *IEEE Access*, vol. 8, p. 63885–63900, 2020; https://doi.org/10.1109/ACCESS.2020.2985258
13. A. Jain, V. Karthikeyan, B. Sahana, B. R. Shambhavi, K. Sindhu and S. Balaji, "Demand Forecasting for E-Commerce Platforms", in 2020 *IEEE International Conference for Innovation in Technology* (INOCON), 2020; https://doi.org/10.1109/INOCON50539.2020.9298395
14. K. Deepa and G. Raghuram, "Sales Forecasting Using Machine Learning Models", *Annals of the Romanian Society for Cell Biology*, p. 3928–3936, 2021; https://www.annalsofrscb.ro/index.php/journal/article/view/5059
15. A. Tony, P. Kumar and S. Rohith Jefferson, "A Study of Demand and Sales Forecasting Model using Machine Learning Algorithm", *Psychology and Education Journal*, vol. 58, p. 10182–10194, 2021; https://doi.org/10.17762/pae.v58i2.3988