Time series modeling for beer sales prediction using the SARIMAX method

Diego Menezes Campos[1]*; Bruno Bourgard Magalhães Garcia [2]

[1] Food and beverage company. Marketing Analyst III - Performance. Rua Pereira Valente, 595, apt. 802 – Meireles; 60160-250 Fortaleza, CE, Brazil
[2] PhD student in Computer Science at the State University of Rio de Janeiro. Rua São Francisco Xavier, 524 – Maracanã; 20.550-900 Rio de Janeiro, RJ, Brazil.
* corresponding author : diegomcamp@hotmail.com

Time series modeling for beer sales prediction using the SARIMAX method

Summary

In the context of companies that work with sales of consumer goods, an accurate estimate of expected sales is essential. This not only aligns expectations and helps in the creation of commercial goals, but also plays a fundamental role in the projection of demand, which can impact an entire production chain of inputs and man-hours. Thus, the objective of this work was to contribute to the prediction of beer demand in a company in the food industry, in order to improve the model previously used. This prediction was made through the multivariate time series model known as SARIMAX (“ Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors ”), which uses an autoregression technique that takes into account the seasonality of sample data, aligned with exogenous predictor variables. This model was chosen due to its robustness and wide use in seasonal forecasts. The application of SARIMAX resulted in a significant improvement in the accuracy of estimates for the states of Ceará and Bahia in the year 2023, presenting error metrics — “ Mean Absolute Percentual Error [MAPE]”, “Root Mean Square Error [RMSE]” and “ Mean Absolute Error [MAE]” — substantially better compared to previously employed methods based on the first 4 months of the year 2023. Consequently, the developed model outperformed the previous one, which was based on a platform called KNIME, with error metrics returning values on average 74.3% lower.

Keywords : Demand analysis; Beverage sector; Supervised models; Sales forecasting .

Abstract

Time series modeling for beer sales forecasting using the SARIMAX method

In the context of companies dealing with consumer goods sales, accurate sales forecasting is imperative. This not only aligns expectations and aids in setting commercial targets, but it also plays a crucial role in demand projection, which can impact an entire supply chain of materials and labor hours. Therefore, the objective of this thesis was to contribute to the beer demand prediction for a company in the foods and beverages industry, aiming to improve the previously used model. This prediction was performed using the multivariate time-series model known as SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors), which employs an autoregressive technique that considers the seasonality of the sample data, aligned with exogenous predictor variables. This model was chosen for its robustness and widespread use in seasonal forecasts. The application of SARIMAX resulted in a significant improvement in forecast accuracy for the states of Ceará and Bahia in 2023, presenting substantially better error metrics — “Mean Absolute Percentual Error [MAPE]”, “Root Mean Square Error [RMSE]” and “Mean Absolute Error [MAE]” — compared to methods previously employed as of the first 4 months of 2023. called KNIME, with error metrics returning values 74.3% smaller on average.

Keywords: Demand analysis; Beverage sector; Supervised models; Sales forecasting.

Introduction

Since the dawn of humanity, there has always been a fascination with predicting the future. It is common to find, in mythologies, mystical beings who had control over time and destiny, such as the Norns , in Norse mythology, or the sisters of destiny, in Greek mythology. According to (Hale, Boer, Chanton, & Spiller, 2003), in ancient Greece, many people went to the temple of Apollo, in the mountains of Delphi, in order to consult the powerful oracle who, according to reports, received divine inspiration by inhaling the gases coming from an opening in the ground.

The tradition of relying on prophecies has permeated human history up until the present day, but with the advent of the scientific method, the ability to “predict the future” became much more tangible. As businesses saw the need to scale their production based on estimates of future sales, various methods of “ sales forecasting ” were developed. According to (French, 2016), “ sales forecasting ” (or sales forecasting) is the use of past sales information to predict short- or long-term performance in order to enable adequate financial planning. As reported (Mills, 2019), sales forecasts make a big difference within companies, allowing for more efficient operational planning, better targeting of advertising campaigns and adjustment of production levels that allows for minimal losses related to lack of stock or “ shelf life ”. life ” (term designated for products that are beyond the minimum commercial expiration date negotiated with the buyer).

In this context, we will examine a medium-sized food and beverage company that operates in the North-Northeast region of the country, which for the purposes of anonymization, will be called company X. In Brazil, the beer sector, until 2018, grew an average of 5% per year since 2004 and was responsible for 1.6% of the National GDP, placing the nation as the third largest beer producer in the world (Diniz, Nunes, Rosa, & Calife, 2019). Nevertheless, pre -pandemic growth is slowly returning, with forecasts made by ETENE (Technical Office of Economic Studies of the Northeast) of a resumption of growth of 5.4% per year until 2025. (Viana, 2022). Historically, the beer industry is considered extremely seasonal (Gevorgyan, 2019)and its sales modeling can take into account several factors, such as the macroeconomic scenario, climate fluctuations and holidays. (Hirche, Haensch, & Lockshin, 2021). Therefore, a good forecast needs to take into account a series of variables in addition to those mentioned above, which makes the forecast model quite complex.

At company X, the sales forecast takes the form of a demand forecast that is passed on to the factories so they can plan the following year's production. This forecast is generated in December of the previous year and serves as the basis for the contractual target called "BP Target", which becomes a volume floor for the sales team. The BP target is reviewed monthly

taking into account the current market scenario and the agreement with the bottling plant undergoes changes to avoid disputes, thus there is fluctuation in the model used for the prediction. In the beverage sector, the two main divisions are between alcoholic and non-alcoholic. Historically, predictions for the non-alcoholic sector are more accurate, and this is due to two distinct factors: the first is that juices, soft drinks, energy drinks and the like have a more stable demand, with a more pronounced seasonality (Lin & Hsu, 2002), while alcoholic drinks in general (mainly beers, the focus of this work) suffer from a much more aggressive variation in consumption and brand hegemony, so that an intuitive examination of their sales graph does not show notable seasonality and their volume variation in the year is often not uniform when compared to previous years (Diniz, Nunes, Rosa, & Calife, 2019).

The second factor is that the alcohol demand forecast is done using a software called KNIME. According to the program's own website , KNIME is an “open (KNIME, 2023)source ” analysis platform . with an intuitive interface that allows the user to perform analyses of any level of complexity – from spreadsheet automation to ETL and “ machine learning ”. It works through a “workflow” system in which several “nodes” (operational boxes) are interconnected, creating an automatic process on top of an original file, such as an Excel data spreadsheet or a .json file.

The non-alcoholic sector uses R software and various machine learning methods. (ARIMA, SARIMA, Neural Networks, Regression, etc.) to select the model with the best accuracy. Thus, the focus of this project will be to apply a unique method of “ sales forcasting ” in the beer sector that has greater accuracy than the “output” company's current status through KNIME.

Material and Methods

The main method utilized was a sales forecasting technique known as SARIMAX (“ seasonal autoregressive integrated moving average with exogenous regressors”, or autoregressive seasonal integrated moving average with exogenous regressors). The SARIMAX model is, in fact, a special case of the ARIMA model (“autoregressive integrated moving average”) in which there is seasonality in the data analyzed. Since beer sales are seasonal and, therefore, are affected by different events and commemorative dates throughout the year, the modeling followed this path. The “x” at the end refers to the presence of exogenous variables, that is, variables whose measurements are determined outside the model and are imposed within the model as its inputs (Mankiw, 2018). These variables became necessary due to the need to particularize the beer model for the case of company X and

provide greater prediction accuracy satisfactory for the exacerbated fluctuation of sales (Gevorgyan, 2019).

The programming language chosen to implement the method was Python, version 3.10.7, through the Visual Studio Code compiler , both due to the author's familiarity and the multitude of Python examples and applications available online. To develop the code, some of the most robust and popular libraries were used: pandas (creation and manipulation of dataframes, generation of graphs, format conversion, etc.), numpy (arrays for comparison of results), statsmodels (creation and fitting of the model) and matplotlib (plotting graphs for comparative purposes). The pipeline used to carry out the study is seen in Figure 1:

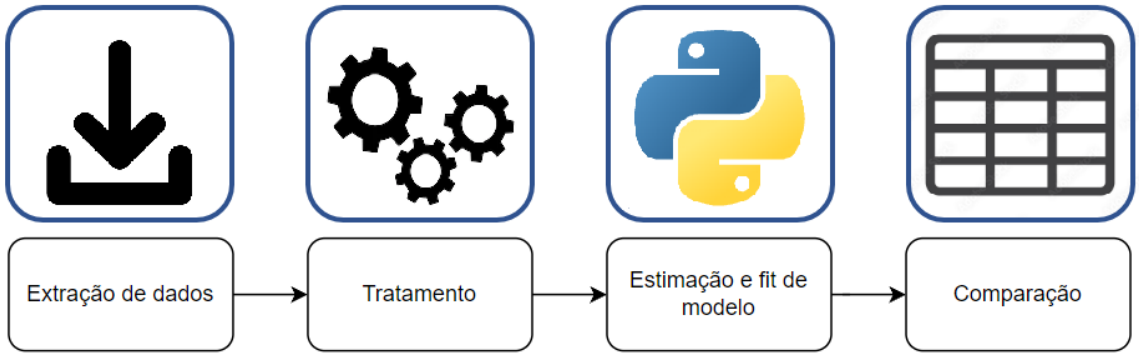

Figure 1. Algorithm development pipeline
Source: Original research data

All volume data was extracted directly from the company's sales system (“Salient Interactive Miner”) in Hectoliters, but underwent a mathematical operation when displaying them in order to preserve sensitive data. This data was then processed via Microsoft Excel to be in a layout suitable for use within Python. The predictor variable (volume) was ordered in months, from June 2018 to December 2022 (the date on which the first edition of Meta BP would be sent for analysis by the board), with the other columns representing the exogenous variables — Broad National Consumer Price Index [IPCA], Monthly Trade Survey [PMC] TOTAL (published by IBGE monthly), PMC for hypermarkets, supermarkets, food and beverages [PMC HS A&B], unemployment (refers to the unemployment rate) and aid (binary variable with 0 for no aid payment during the pandemic and 1 for aid payment). — necessary to create the model.

The initial ARIMA model was estimated manually in order to better understand the behavior of the volume and how it presents itself under the different parameters. Due to the scope of this work, it was decided to use only data from the states of Ceará and Bahia. The state of Ceará served as an initial proof of concept and Bahia served as a direct application of

the developed methodology, with expectations of, at a later date, applying the model to the other UFs .

According to (Hyndman & Athanasopoulos, 2013), the ARIMA and exponential smoothing models are among the most widely used for prediction in time series. However, due to the complex nature of the data in question, the need to work with exogenous variables makes the use of the SARIMAX model much more advantageous than the exponential smoothing model with exogenous variables (ESTX), which has little applicability and few studies on the subject. The ARIMA model is an autoregressive model, that is, the variable of interest is projected using previous values of itself. However, it can only be used in stationary data series. If $\{y_t\}$it is a stationary series, then for all *s ,* the distribution of $(y_t, \ldots, y_{t+s})$does not depend on t. Our data do not respect this rule, and to make them stationary, it was necessary to differentiate them, that is, compute the differences between sequential observations (Hyndman & Athanasopoulos, 2013).

The ARIMA model has three distinct parameters: p, d and q, which are fundamental components for modeling the time series:

- The parameter p refers to the order of the autoregressive (AR) component of the model. Specifically, p denotes the number of " lags ," or past observations, that will be included as predictors in the model. In other words, p is the amount of direct autocorrelation that the model will account for.
- The parameter d refers to the degree of differencing (I) in the time series. Differencing is a method used to make a time series stationary by subtracting the current observation from an observation at a previous time interval. Specifically, d represents the number of times the time series needs to be differencing to achieve stationarity .
- The q parameter is the order of the moving average (MA) component of the model. The moving average component refers to the model error as a linear combination of the past error terms. The q parameter therefore represents the number of past error terms that the model will consider. (Hyndman & Athanasopoulos, 2013)

Together, these three parameters describe how the ARIMA model will consider autocorrelation in the time series, how it will address stationarity of the series, and how it will model forecast errors. (Prabhakaran, 2021)

With the help of some functions from the statsmodels library, such as ndiffs and plot_acf, it was possible to plot the original series, its differentiations and the respective autocorrelations. From the analysis of these graphs, an optimal configuration of p, d and q was found and the model fit was made. Since the model initially worked on was an ARIMA, the estimation of p, d and q was only useful because, later, in SARIMA, these parameters are needed. The ARIMA model obtained initially served only to project trends and did not satisfy the initial proposal. To take into account the seasonal fluctuations present in the data, it was

necessary to fit the model into the SARIMA format. A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA model that we have seen so far. It is written according to the form (1):

$$SARIMA = (p, d, q)x(P, D, Q)_m \quad (1)$$

Where: p, d and q are the parameters explained previously, P, D and Q are the same parameters for the seasonal component and m represents the seasonal period (e.g. the number of observations in a year). We write in capital letters the seasonal parts of the model and in lowercase letters the non-seasonal parts. The seasonal part of the model consists of terms that are similar to the non-seasonal components, but involve backwards from the seasonal period. For example, an ARIMA(1,1,1)(1,1,1) $_4$ is for quarterly data (m = 4).

The additional seasonal terms are simply multiplied by the non-seasonal terms. In Python, this method was implemented using the auto_arima function from the pmdarima library, with the seasonal = true component . This function compares all possible combinations of (p, d, q)x(P, D, Q) and provides the " summary " *of the configuration with the lowest* Akaike Information Criterion [AIC] (a statistical tool that simultaneously assesses the complexity of a model and its predictive power). Although the metric is useful, it was only used for parameter selection, but not for comparing models per se, given that, by taking into account the complexity of the model, the metric is not useful if the main interest is the accuracy of predictions.

With the SARIMA model in hand, the next step was to define the exogenous variables. According to the authors of (Yves R. Sagaert, 2017), There is an undeniable impact of macroeconomic variables on sales in a wide range of economic sectors, thus highlighting the imperative of their incorporation. In the non-alcoholic sector, some variables used were macroeconomic indicators that are measured monthly and quarterly, such as IPCA, PMC TOTAL, PMC HS A&B, unemployment and aid. Thus, these variables were chosen because, in addition to having intrinsic relevance, they were proven to be effective within the company in predicting beverage sales. The date column was divided by month, so that all of these indicators required monthly values for use.

From March 2020 onwards, the Covid-19 pandemic radically changed the way people consume in the market, given the series of restrictions on the operation of several establishments. However, a COVID-19 variable was not created to represent the effects of the pandemic on alcohol consumption because, according to the OPAN report on alcohol consumption during the pandemic, consumption levels increased slightly but remained similar to pre -pandemic periods (Organização Pan-Americana da Saúde, 2020).

Company X has partnerships with different beer brands, acting more as a distributor and bottler than a producer. These beer brands, in turn, have different contracts with other bottlers in other parts of Brazil, so that Company X's beer portfolio varies from state to state (for example, brand A is distributed by Company X in Ceará, but not in Acre, where it is distributed by Company Y). Therefore, it was necessary to create “dummy ” variables (variable created to incorporate categorical categories, which, by nature, are qualitative, in models that require a numerical input) for the presence of brands in the portfolio in the state, since this presence changed over time, with different brands entering and leaving the company's distribution portfolio in the state over the years of analysis. The “ dummy ” variables functioned as 1 for presence of sales of the brand and 0 for non-presence.

It is known that the sale of alcoholic beverages and the weather are closely related, as is the occurrence of holidays. In (Hirche, Haensch, & Lockshin, 2021), the authors set out to quantify this relationship in different states of the USA, concluding that, with a 1°C increase in temperature, the average store sells an additional 11.4 liters of beer, 4.6 liters of liquor and 0.3 liters of white wine per week. Similarly, in a week with a holiday, the average store would sell an additional 72 liters of beer, 10 liters of liquor and 9 liters of white wine. In order to corroborate the authors' research and adopt new exogenous variables, two variables were created: one, called "Temperature", presented the monthly average temperature of the state of Ceará (Instituto Nacional de Meteorologia, 2023), and the other, called "Carnival", presented whether or not Carnival occurred in that month (therefore, it is a binary variable).

In order to avoid negative predicted volume values, the model was created on top of the natural logarithm of the volume and, later, converted back to real values through an exponential on the values with fit . After this transformation, the predicted values were directly compared with the real values of the first four months of the year 2023 through three metrics:

“Mean Absolute Percentage Error” [MAPE] : This is a statistical measure that expresses the magnitude of the error of a forecast model as a percentage of the actual observed values. By dividing the sum of the absolute value of the differences between the forecasts and the actual values by the total number of observations, it is possible to determine the average of the forecast error ratio. The percentage nature of MAPE makes this metric particularly useful for comparisons between time series of different scales. In general, a lower MAPE indicates a better forecast model, with less forecast error.

Root Mean Square Error [RMSE] : RMSE is a measure of accuracy that is calculated as the square root of the mean of the squared errors. It is a useful indicator of the magnitude of prediction errors, with a greater emphasis on larger errors due to its quadratic nature. RMSE has the advantage of being in the same units as the dependent variable, which makes it easier

to interpret the results. Models with lower RMSE are preferred, as it suggests that the model has smaller prediction errors.

Mean Absolute Error [MAE] : MAE is a prediction error metric that is calculated as the mean of the absolute values of the differences between the predictions and the observed values. MAE is a useful metric because it is easy to understand, is in the same units as the output variable, and gives an idea of the size of the mean prediction error. However, because it does not penalize large errors as much as RMSE, it may not be as sensitive to outliers in the data. A model with a smaller MAE is preferable, as it indicates that the model has smaller prediction errors.

Due to the chaotic nature of beer volume, a 90% confidence interval was adopted for the entire selection of values.

Results and discussion

Ceara

For the first run of the model, with all the exogenous variables mentioned, the results in Figure 2 were obtained:

SARIMAX Results

Dep. Variable:	Volume	No. Observations:	55
Model:	SARIMAX(0, 1, 0, 12)	Log Likelihood	-7.545
Date:	Thu, 01 Jun 2023	AIC	47.089
Time:	21:10:31	BIC	74.892
Sample:	06-30-2018 - 12-31-2022	HQIC	57.280
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
Dummy_A	-0.0941	0.170	-0.555	0.579	-0.426	0.238
Dummy_B	-0.0632	0.170	-0.373	0.709	-0.396	0.269
Dummy_E	-0.1349	0.117	-1.152	0.249	-0.364	0.095
Dummy_E	0.0939	0.170	0.554	0.580	-0.238	0.426
Dummy_H	0.6770	0.450	1.505	0.132	-0.205	1.559
Dummy_K	-3.24e-10	nan	nan	nan	nan	nan
Dummy_S	-4.611e-10	nan	nan	nan	nan	nan
Dummy_T	-0.1349	0.117	-1.152	0.249	-0.364	0.095
Desocupação	7.8222	4.160	1.880	0.060	-0.332	15.976
IPCA	-3.1189	3.872	-0.805	0.421	-10.708	4.470
PMC H S A&B	-0.0286	0.015	-1.946	0.052	-0.057	0.000
PMC TOTAL	0.0100	0.005	1.967	0.049	3.36e-05	0.020
Auxílio	0.0539	0.153	0.351	0.725	-0.247	0.355
Carnaval	0.3718	0.213	1.750	0.080	-0.045	0.788
Temperatura	0.1063	0.183	0.582	0.561	-0.252	0.464
sigma2	0.0838	0.026	3.243	0.001	0.033	0.134

Ljung-Box (L1) (Q):	1.71	Jarque-Bera (JB):	0.29
Prob(Q):	0.19	Prob(JB):	0.86
Heteroskedasticity (H):	1.07	Skew:	-0.03
Prob(H) (two-sided):	0.91	Kurtosis:	2.60

Figure 2. Summary of SARIMAX in Ceará with all exogenous variables
Source: Original research results

In the initial model, the main intention was to select which variables are significant for prediction, at a confidence level of 90%, i.e., with p value < 0.1. According to the criteria adopted in this article, only PMC HS A&B (0.052), PMC TOTAL (0.049), Carnaval (0.080) and Desocupação (0.060) passed the selection. The other variables did not present a significant contribution, with two of them without p value (this is due to the fact that they were constant throughout the period evaluated, i.e., they did not contribute to the variation in results). One explanation for the exit and entry of beers not being an adequate predictor variable is the fact that, after the exit of a brand, another brand in the same beer category (for example, “premium”, which is a category composed of high-cost and refined beers, or “economy”, a category of cheaper beers) “captures” the “gap” created by the previous one. Similarly, the entry of a brand, as it is unable to capture such a large “share”, means that the pre-existing volume is diluted and the results do not change so significantly.

Regarding the results of the temperature variable, (Hirche, Haensch, & Lockshin, 2021) they state that, for intra- US results, the effect of temperature variation was less visible in warmer climates. Since negative temperature variation did not have as much influence, a positive variation is less common or less noticeable in warm climates. Extrapolating this idea to the Brazilian reality, the state examined has a predominantly warm climate, which takes away the predictive power of the temperature variable. However, the carnival variable passed the initial selection, which gives some credibility to the article.

The null hypothesis of heteroscedasticity (non-constant error variability) was rejected, i.e., the model is homoscedastic , and, based on the Skew and Kurtosis values , the residuals have a distribution close to normality, which gives a positive character to its predictive capacity. When running this model, Python issued a warning that it failed in the maximum loglik optimization, which may suggest, among other things, multicollinearity (strong linear relationship between independent variables) between variables. Calculating the Variance Inflation Factor [VIF] (indicative of multicollinearity) for the PMC variables, Table 1 was obtained:

Table 1 – VIF results for Ceará model

Variable	VIF
TOTAL PMC	1.551898
PMC HS A&B	1.551898
Intercept	198.38

Source: Original research results

Despite the alarming VIF with the intercept, the individual VIFs were within the expected normal range (up to 5 is a usual parameter, according to (Hyndman & Athanasopoulos, 2013)).

With this, the models were ready to be tested. Two methodologies for extrapolating exogenous variables were used: multivariate linear regression and SARIMA. For the period of the first four months of 2023, the selected exogenous variables PMC HS A&B, PMC TOTAL and Unemployment were extrapolated and tested. When compared, extrapolation values from linear regression provided lower values for the test metrics MAPE, RMSE and MAE than values from SARIMA, therefore, the extrapolated values followed that method.

A series of models were tested sequentially in order to choose the best combination arrangement among the predictor variables. A permutation of the four variables was tested between several iterations of SARIMAX, always comparing the absolute values of MAPE, RMSE and MAE. After several simulations and tests with one, two, three and four variables, a specific SARIMAX model stood out against the others and presented the best result when compared to the others, which would be SARIMAX with only one exogenous variable: carnival.

When isolated from each other, the macroeconomic variables always ended up with p values much higher than expected, which caused the tested models to present results inferior to those obtained by a simple SARIMA model. When simulating SARIMAX with the carnival variable, the summary in Figure 3 was obtained:

SARIMAX Results

Dep. Variable:	Volume	No. Observations:	55
Model:	SARIMAX(0, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-18.151
Date:	Sat, 03 Jun 2023	AIC	44.301
Time:	01:52:05	BIC	51.252
Sample:	06-30-2018	HQIC	46.849
	- 12-31-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
Carnaval	0.3761	0.209	1.799	0.072	-0.034	0.786
ma.L1	-0.4491	0.157	-2.858	0.004	-0.757	-0.141
ma.S.L12	-0.6369	0.431	-1.477	0.140	-1.482	0.208
sigma2	0.1198	0.041	2.945	0.003	0.040	0.200

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	1.16
Prob(Q):	0.81	Prob(JB):	0.56
Heteroskedasticity (H):	1.32	Skew:	-0.17
Prob(H) (two-sided):	0.61	Kurtosis:	2.26

Figure 3. Summary of SARIMAX in Ceará with only carnival as an exogenous variable
Source: Original research results

In the chosen model, Carnival and ar.L 1 have a p value less than 0.1, i.e., we reject the null hypothesis that the coefficient is equal to zero and accept that these predictor variables have a significant effect on the dependent variable (with a 10% confidence interval). This

demonstrates that the seasonality associated with Carnival appears significantly in the data and that the lag1 autocorrelations are well captured. Since Prob (Q) > 0.05, we cannot reject the null hypothesis that the residuals are not correlated, which suggests that the model adequately captured the temporal dependence in the data and that the residuals resemble white noise - i.e., the residuals are basically a series of random and independent values.

Furthermore, the model is homoscedastic (constant variance at all levels of the independent variables) and its residuals are normally distributed (Prob (JB) > 0.05 and good values of " Skew " and " Kurtosis "). When we compare AIC and "Log Likelihood " (the logarithm of the likelihood, which quantifies how well the model fits the observed data, with higher values indicating a better fit) with the initial model, we notice a small worsening, but this is most likely due to "overfitting" (biased adherence of the model's prediction to the real data) of the initial model and a change in seasonality for the year 2023, which began with a rise in the beer sector below that expected by experts (Guia da Cerveja, 2023).

The SARIMAX model in question was then compared with two other models: the model already used by company X, called the "Goals Model", and a standard SARIMA model, without any exogenous variables. After comparing these three models with the actual sales data, Table 2 was obtained:

Table 2 – MAPE, RMSE and MAE for Ceará

Model	MAPE	RMSE	MAE
Goals Model	39.1	2392.1	2048.6
SARIMA Model	26.1	1367.4	1322.1
SARIMAX Model	15.5	926.2	752.6

Source: Original research results

According to the table, the chosen SARIMAX model was better than the other two models, in all requirements: the three parameters presented better results when compared to the values of the year 2023, with an average error reduction of 61.6%. The reason for this improvement can be seen in Figure 4:

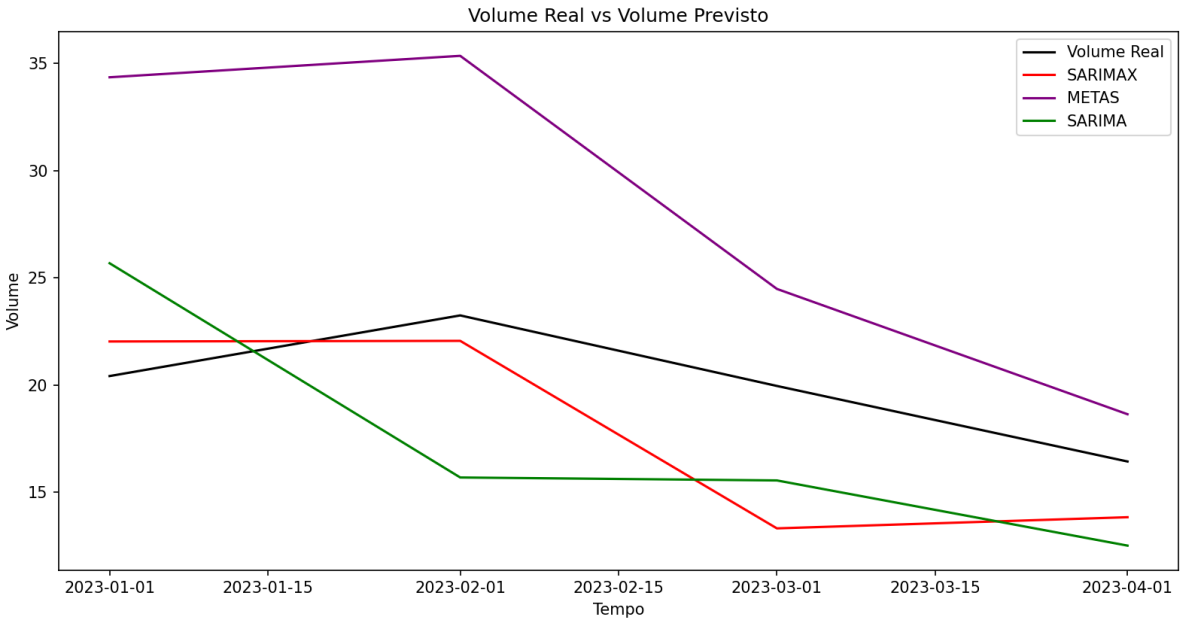

Figure 4. Volume comparison in the three models for Ceará

Source: Original research results

Note: * the data underwent a mathematical operation in order to preserve the anonymity of the volumes in Hectoliters of the Ceará region

Carnival 2023 took place at the end of February. While the SARIMA model predicted a sharp drop in volume in that month and did not take into account the sales boom usually caused by the event (which can occur in either February or March, depending on the year), the model with the “Carnival” variable predicted a higher volume in the month, thus increasing its accuracy. The reason why the two models do not agree before and after Carnival is that, for the entry of the exogenous variable, an optimal configuration of (p, d, q) x (P, D, Q) had to be found again, which changes the behavior of the model in the other periods of the year.

Bahia

To estimate the Bahia model, some generalizations learned in the estimation of the Ceará model were necessary. No “transfer learning” techniques were used as these are commonly associated with “deep learning” models, especially in scenarios where there is a large amount of data for the base model (such as image recognition) and little data for the required model (ZHUANG, et al., 2020). In addition, Bahia has a larger population, a diverse climate and distinct festivities that may not be present or as prominent in Ceará, which further contributes to the biases and specific peculiarities of the previously developed model not being repeated in the new state.

After identifying the indexes (p,d ,q)x(P,D,Q), the model with the exogenous variable carnival was created and its summary was shown in Figure 4:

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                            Volume   No. Observations:                   54
Model:             SARIMAX(1, 2, 0)x(1, 1, 0, 12)   Log Likelihood                -13.211
Date:                            Mon, 18 Sep 2023   AIC                             34.422
Time:                                    22:27:54   BIC                             39.606
Sample:                                07-31-2018   HQIC                            35.964
                                     - 12-31-2022
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Carnaval       0.2959      0.154      1.918      0.055      -0.007       0.598
ar.L1         -0.4944      0.169     -2.917      0.004      -0.827      -0.162
ar.S.L12      -0.6326      0.235     -2.696      0.007      -1.092      -0.173
sigma2         0.1558      0.046      3.388      0.001       0.066       0.246
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                 1.73
Prob(Q):                              0.82   Prob(JB):                         0.42
Heteroskedasticity (H):               3.26   Skew:                            -0.59
Prob(H) (two-sided):                  0.09   Kurtosis:                         3.37
===================================================================================
```

Figure 4. Summary of SARIMAX in Bahia with only Carnival as an exogenous variable
Source: Original research results

The p-values of the predictor variables in the SARIMAX model are in accordance with the adopted significance threshold of 10%, corroborating their relevance in volume prediction. The Ljung -Box test (a statistical test to verify autocorrelation in the residuals of a time series model), with a Q-value of 0.05 and an associated p-value of 0.82, suggests that the model residuals did not present significant autocorrelations until the first “lag”. This reinforces the model's ability to capture the temporal dependence structure in the data. Additionally, the heteroscedasticity test, with a p-value of 0.09, indicates that the residuals were homoscedastic. By submitting the model to the MAPE, MAE and RMSE indicators and comparing these results with the proposed volume target values for the first four months of the year, Table 3 was obtained:

Table 3 – MAPE, RMSE and MAE for Bahia

Model	MAPE	RMSE	MAE
Goals Model	67.7	2537.7	2413.3
SARIMAX Model	8.4	360.5	303.1

Source: Original research results

The results in Table 3 showed that the developed SARIMAX model performed better than the previously used model and was able to explain the behavior of the volume time series analyzed with greater precision, obtaining an average error reduction of 86.9%. This improvement was clear when analyzing the plot of the first 4 months of the time series vs. the target, shown in Figure 5:

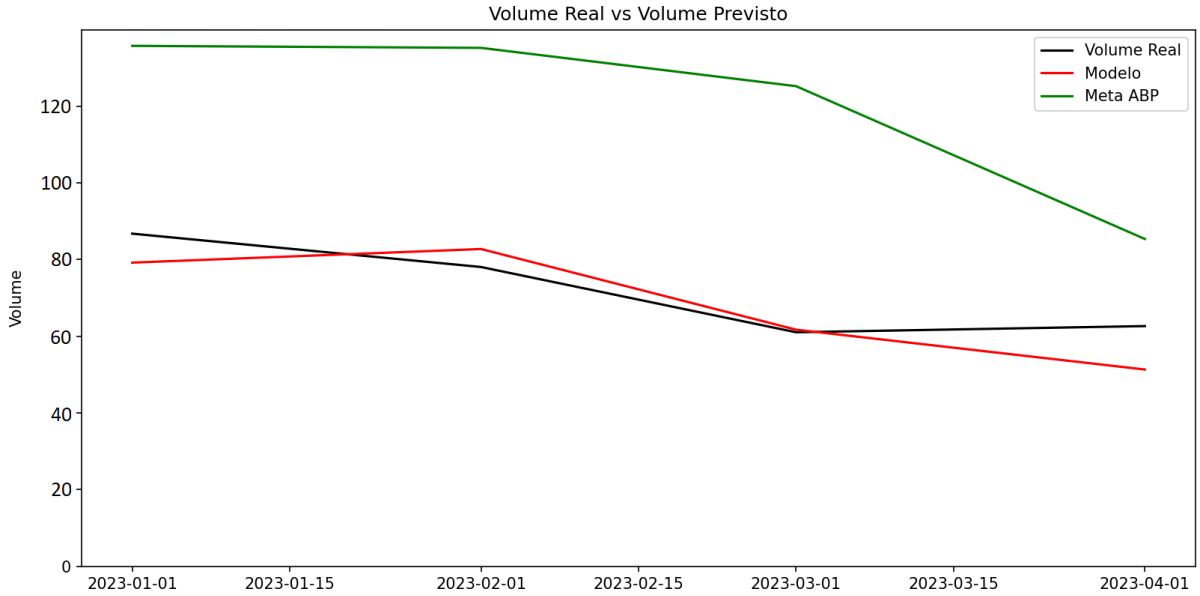

Figure 5. SARIMAX model volume comparison for Bahia
Source: Original research results
Note: * the data underwent a mathematical operation in order to preserve the anonymity of the volumes in Hectoliters of the Bahia region

In addition to better predicting volume results, the model also provided some interesting insights into the state's behavior. Carnival, which took place in February, did not increase beer volume as observed in previous years. In addition, the subsequent upward trend in April was due to a commercial incentive that came into effect the following month and was not adequately modeled, even providing an opportunity for improvement in the SARIMAX used.

Final Considerations

Thus, a model was established that outperformed the previous model adopted by the company, achieving an average error reduction of 74.3% between the two states. Through a rigorous process of time series modeling with exogenous variables, it was possible to develop the proof of concept in the state of Ceará and successfully apply it in the state of Bahia, obtaining good results in both attempts. Since the planned result of reducing error metrics was achieved, it will be possible to expand the method to all the states served and use the results as a basis for creating volume targets for the year 2024. Possible improvements and future

studies may come from the inclusion of more exogenous variables and larger tests with different p, q, and d indices.

Thanks

I would like to thank my mother and grandmother, two essential people in my life. I would also like to thank my friends, my advisor and my girlfriend, who were fundamental pillars in this difficult journey.

References

Diniz, I. m., Nunes, DM, Rosa, VA, & Calife , NF 2019. APPLICATION OF A DEMAND FORECASTING MODEL IN A BEER COMPANY. In: Brazilian Journal of Production Engineering. Proceedings… p. 120-138.

French, J. 2016. Economic determinants of wine consumption in Thailand. In: International journal of economics and business research. Annals… p. 334-347.

Gevorgyan, R. 2019. Statistical analysis of time series and investigation of seasonal fluctuations on beer production. Proceedings of the 3rd International Conference on Business and Information Management. Annals… p. 134-137.

Beer Guide. Alcoholic beverage production begins 2023 with a 1.4% increase. Available at: < https://guiadacervejabr.com/producao-bebidas-alcoolicas-2023-janeiro/ >. Accessed on: April 3 , 2023.

Hale, JR, Boer, JZ, Chanton, JP, & Spiller, HA 2003. Questioning the Delphic Oracle. Scientific American, p. 66-73.

Hirche , M., Haensch, J., & Lockshin , L. 2021. Comparing the daytime temperature and holiday effects on retail sales of alcoholic beverages – a time-series analysis. International Journal of Wine Business Research. Annals… p. 432-455.

Hyndman, RJ, & Athanasopoulos, G. 2013. Forecasting: principles and practice (Vol. III). OTexts .

National Institute of Meteorology. Annual historical data. Available at: < https://portal.inmet.gov.br/dadoshistoricos >. Accessed on: May 25 , 2023.

KNIME. (2023, May). Knime Software Overview. Available at: < https://www.knime.com/software-overview >. Accessed on: May 24 , 2023.

Lin, C.-T., & Hsu , P.-F. 2002. Forecast of Non-alcoholic Beverage Sales in Taiwan Using the Gray Theory. Asia Pacific Journal of Marketing and Logistics. Annals… p. 3-12.

Mankiw, N. G. 2018. Macroeconomics. 10th ed. Worth Publishers.

Mills, T. C. 2019. Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting. 1st ed. Academic Press.

Pan American Health Organization. 2020. Alcohol use during the COVID-19 pandemic in Latin America and the Caribbean. World Health Organization.

Prabhakaran, S. 2021. ARIMA Model – Complete Guide to Time Series Forecasting in Python. Available at: < https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ >. Accessed on: July 16, 2023.

Sagaert , Y. R., Aghezzaf , E.-H., Kourentzes , N., & Desmet , B. (2017). Tactical sales forecasting using a very large set of macroeconomic indicators. European Journal of Operational Research . Anais... p. 558 – 569.

Viana, FL 2022. ETENE sector notebook: Alcoholic beverage industry. Available at: < https://www.bnb.gov.br/s482-dspace/bitstream/123456789/1159/3/2022_CDS_216.pdf >. Accessed on: July 5 , 2023

ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., . . . HE, Q. 2020. A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE. Anais... p. 1-34.