University of Texas at Dallas (2023), 1–4

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE TYPE

Store Sales Forecasting using Time Series methods

Harikrishna Dev,[1] Amit Mishra,[2] Jaswanth Reddy Yanumula,[3] Abhiram Mane,[4] and Soumyajit Mishra[5]

[1]HXD220000, The University of Texas at Dallas, Dallas, 75080, TX, Richardson
[2]AKM220000, The University of Texas at Dallas, Dallas, 75080, TX, Richardson
[3]JXY210004, The University of Texas at Dallas, Dallas, 75080, TX, Richardson
[4]AXM220137, The University of Texas at Dallas, Dallas, 75080, TX, Richardson
[5]SXM220133, The University of Texas at Dallas, Dallas, 75080, TX, Richardson
Author for correspondence: H. Dev, Email: harikrishna.dev@utdallas.edu.

Abstract

This paper examines the impact of gas prices on the economy and daily wages of the population in Ecuador, a country known for its reliable oil resources. The analysis focuses on Favorita Corporation, Ecuador's favorite retailer, which adapts its product offerings and experiences to the local realities and customers' needs. The study draws on data from 2016, which shows that Ecuador holds a significant amount of proven oil reserves, ranking 19th in the world. The findings shed light on the ways in which gas prices affect Ecuador's economy and the implications for businesses like Favorita Corporation that rely on consumer spending. Overall, this paper contributes to a better understanding of the complex relationship between energy resources, business operations, and the broader socio-economic context in Ecuador.

Keywords: Store Sales Forecasting, Time series, Linear regression

1. Literature Review

Time Series Analysis and Forecasting is a statistical technique used to analyze and predict patterns in time-based data considering the data points over a period. This helps us understand the design and underlying data structure from which we can make sense of a trend being followed. As the subject says, "Time Series Analysis," the data is collected based on the individual's or model's requirement in terms of daily, weekly, monthly, or yearly intervals. We evaluate a few components of TSA, like trend, seasonality, cyclicality, and random variation in data, which must be identified. After identifying the details, we create models to help us predict the target variable's future values (forecast). This has a wide range of applications, such as financial forecasting, demand forecasting, weather forecasting, and many more. For this, TSA has a few techniques, namely MA (Moving Average), Exponential smoothing, ARIMA (Autoregressive Integrated Moving Average), and a few Machine Learning Algorithms such as Neural Networks (CNN, RNN, etc.). This choice of technique used purely depends on the type of problem being addressed.

This paper talks about "Demand Forecasting," which means that we here estimate and forecast the future demand for a product/service. This is a part of Supply Chain and Operations Management, where the scope of prediction has a wide range of variables that can be considered. This can be done for various time horizons (short-term, medium-term, and long-term). Methods used for Demand forecasting can either be Qualitative or Quantitative. This paper has a quantitative approach as we use historical data to forecast the future demand for the product. Demand forecasting is essential for businesses to stay competitive and profitable.

We take our reference from Intelligent Demand Forecasting published by Nimai Chand Das Adhikari, Nishanth Domakonda, Chinmaya Chandan, Gaurav Gupta, Rajat Garg, S Teja, Lalit Das, Dr. Ashutosh Misra, where they use the SES model, Moving Average, Croston Model, and Seasonal Linear Regression with Time Series and Regression based algorithms to prove that both algorithms put together to bring out the result close to actual rather than under-forecasting or over-forecasting. Here we can analyze the impact that the model does underfit and overfit when the algorithms are used separately (wrong output). In contrast, when both models are in sync, we get a result much closer to the actual value of the target variable.

Also, a Comparative Study between Classical Methods and Machine Learning Algorithms for TSA published by Heba Salah, Mohammed Hussein, and Ismail Zahran compared Demand Forecasting scope by applying both the algorithms and concluded by giving positives and negatives for Classical Methods and Machine Learning Algorithms. The authors discuss CM and MLA (Back Propagation Neural Network Algorithm). They talk about how variance is terrible for the model and not just one technique gives accurate results for all data points. For the considered dataset, they conclude that CM gives more precise results than MLA as they feel the past trend will further advance in the future than the NN is a black box. CM is better for short-term data than MLA for movement or seasonal data.

2. Data

Our dataset is from Kaggle and is part of the Time series competitions. We chose this dataset as it could have seasonal aspects and significantly impact holidays and the country's economic dependence on oil prices. There are 54 stores and 33 product families in the data. The time series starts on 2013-

01-01 and finishes on 2017-08-31. We also have oil prices for the same period.

The data contains target sales and the time series for the features store nbr, family, and OnPromotion. The store where the goods are sold is identified by the variable store nbr. The family identifies the category of the sold goods. sales provide the overall sales for a family of products at a specific retailer on a particular day. Since products can be sold in fractional units (1.5 kg of cheese, as opposed to 1 bag of chips, for example), fractional values are possible. OnPromotion displays the total number of products from a family that was sold by the retailer on a specific date. Daily oil price. (Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices.) Holidays and Events, with metadata. The transferred column containing holidays that are transferred officially falls on that calendar day but was moved to another date by the government

3. Model Empirical Method

We have evidence of weekly seasonality based on the periodogram and seasonal plot in time series data refers to a pattern of recurring frequency that occurs at fixed intervals of ch as days, weeks, months, or years. These recurring patterns are often associated with seasonal factors, like weather, holidays, or other calendar-related events. They came a significant impact on the behavior of the time series. The periodogram suggests there is a monthly and biweekly factor based on the Store Sales dataset as wages in the public sector are paid out biweekly.

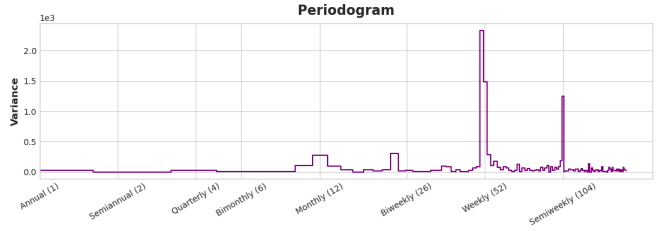

Figure 1. Periodogram with Seasonality

To account for the higher sales pattern over the holidays, we deseasonalize by fitting a regression model to the time series data that includes seasonal dummy variables. The coefficients of the seasonal dummy variables can subsequently be used to adjust the data for seasonality.

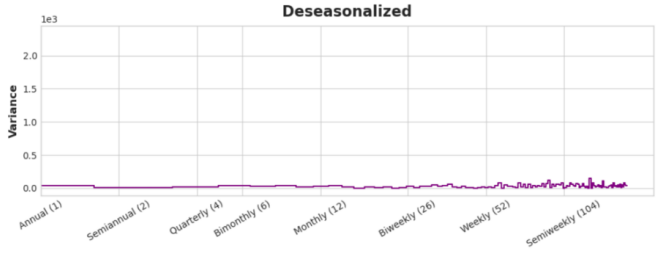

Figure 2. Deseasonalized Periodogram

We can also see the sales peak on Saturday and in the month of December. This could be because 74% is Roman Catholic. This would explain why we observe the majority of the sales being on Saturdays and during the month of December as Christmas is the major festival for the people in Ecuador and Christians tend to visit the church on Sundays while being busy on the weekdays.

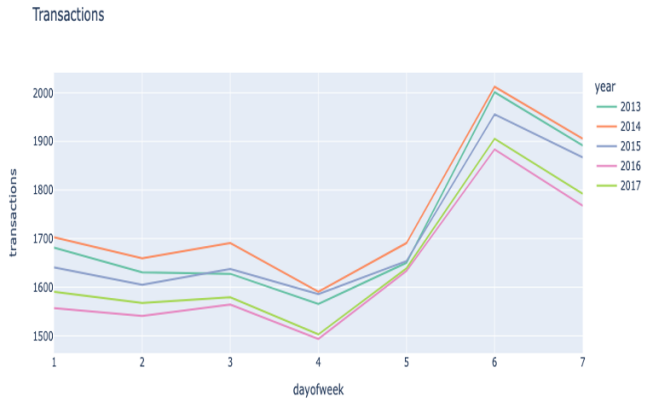

Figure 3. Average Weekly Sales

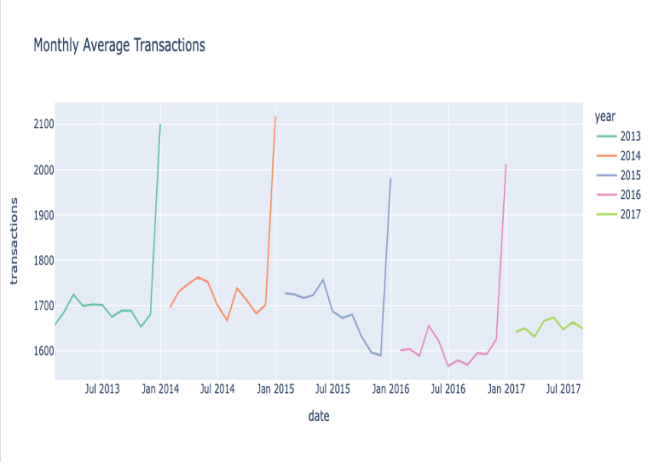

Figure 4. Average Monthly Sales

The economy is one of the biggest problems for governments and people. It affects all things in a good or bad way. In our case, Ecuador is an oil-dependent country. Changing oil prices in Ecuador will cause a variance in the model. We researched Ecuador's economy to be able to understand much better and we found an article from IMF.

Link: https://www.imf.org/en/News/Articles/2019/03/20/NA032119-Ecuador-New-Economic-Plan-Explained

There are some missing data points in the daily oil data as we can see below. Linear Interpolation is suitable for this time series. We can see the trend and predict missing data points when we look at a time series plot of oil prices.

4. Results

We can decompose the data into its components assuming its additive in nature.

Figure 5. Oil prices in USD

$$sales = trend + seasonality + noise$$

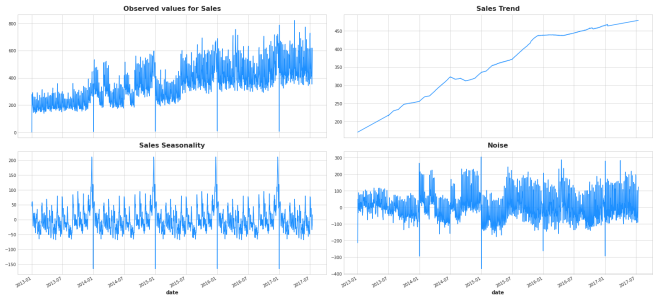

Figure 6. Decomposition of Sales

After running an OLS regression, we were able to get a model which predicts the sales based on the Oil prices, date, and the time of the year.

```
. regress sales dcoilwtico transferred t feb mar apr may jun jul aug sep oct nov dec tue wed thu fri sat sun
```

Source	SS	df	MS	Number of obs	=	1,714
				F(20, 1693)	=	288.34
Model	7.3857e+13	20	3.6928e+12	Prob > F	=	0.0000
Residual	2.1683e+13	1,693	1.2807e+10	R-squared	=	0.7731
				Adj R-squared	=	0.7704
Total	9.5539e+13	1,713	5.5773e+10	Root MSE	=	1.1e+05

sales	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
dcoilwtico	563.7958	180.7797	3.12	0.002	209.2207	918.371
transferred	-165798.9	38114.83	-4.35	0.000	-240556	-91041.74
t	358.318	7.829325	45.77	0.000	342.9618	373.6742
feb	-51383.62	13178.61	-3.90	0.000	-77231.69	-25535.54
mar	-6791.995	12868.52	-0.53	0.598	-32031.88	18447.89
apr	-37668.74	12970.81	-2.90	0.004	-63109.25	-12228.22
may	-43803.17	12849.6	-3.41	0.001	-69005.93	-18600.4
jun	-41264.83	12892.62	-3.20	0.001	-66551.97	-15977.69
jul	-10024.99	12915.69	-0.78	0.438	-35357.38	15307.41
aug	-59886.18	13323.25	-4.49	0.000	-86017.96	-33754.4
sep	10900.69	13828.78	0.79	0.431	-16222.61	38023.98
oct	2953.251	13721.03	0.22	0.830	-23958.72	29865.22
nov	9760.043	13766.8	0.71	0.478	-17241.69	36761.78
dec	153482.8	13641.35	11.25	0.000	126727.1	180238.4
tue	23971.73	10247.45	2.34	0.019	3872.726	44070.74
wed	-60134.03	10216.89	-5.89	0.000	-80173.09	-40094.97
thu	11478.44	10248.78	1.12	0.263	-8623.164	31580.05
fri	210084.3	10548.74	19.92	0.000	189394.3	230774.2
sat	264501.6	10521.43	25.14	0.000	243865.2	285137.9
sun	48132.33	10248.12	4.70	0.000	28032	68232.65
_cons	235207.3	19972.06	11.78	0.000	196034.8	274379.9

Figure 7. STATA regression output

The output shows the results of multiple linear regression with sales as the dependent variable and dcoilwtico, transferred, t, and the months and days of the week as independent variables.

The F-test result indicates that the overall regression model is statistically significant, with a p-value of 0.0000. The adjusted R-squared value is 0.7704, indicating that the model explains 77.04% of the variability in sales after adjusting for the number of independent variables in the model.

We can see that a one-dollar increase in dcoilwtico is associated with a $563.80 increase in sales, while a Tuesday is associated with a $23,971.73 increase in sales, on average.

While conducting Wald tests to check the joint significance of the seasonal dummy variables we get the following results:

```
. test feb mar apr may jun jul aug sep oct nov dec

 ( 1)  feb = 0
 ( 2)  mar = 0
 ( 3)  apr = 0
 ( 4)  may = 0
 ( 5)  jun = 0
 ( 6)  jul = 0
 ( 7)  aug = 0
 ( 8)  sep = 0
 ( 9)  oct = 0
 (10)  nov = 0
 (11)  dec = 0

       F( 11,  1693) =   31.86
            Prob > F =    0.0000
```

Figure 8. Wald test on Month Dummy variables

```
. test tue wed thu fri sat sun

 ( 1)  tue = 0
 ( 2)  wed = 0
 ( 3)  thu = 0
 ( 4)  fri = 0
 ( 5)  sat = 0
 ( 6)  sun = 0

       F(  6,  1693) =  235.39
            Prob > F =    0.0000
```

Figure 9. Wald test on Day Dummy variables

The first test examines whether the coefficients for the months of the year are jointly equal to zero, while the second test examines whether the coefficients for the days of the week are jointly equal to zero. Both tests have very low p-values, indicating that we can reject the null hypothesis that all of the coefficients are equal to zero, and conclude that at least one of the variables has a significant effect on sales.

After predicting the sales for our training data, we predicted future sales and we can see the model takes the seasonality and trend into consideration while forecasting.

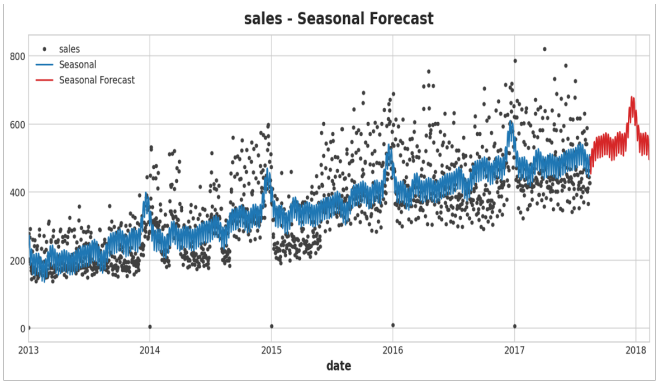

Figure 10. Predicted sales overlayed on Actual sales

We performed Breusch-Pagan/Cook-Weisberg test for heteroskedasticity in a regression analysis and we find that our predicted sales are not homoskedastic in nature. This violates our OLS assumptions.

```
. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of sales

H0: Constant variance

    chi2(1) =  90.18
Prob > chi2 = 0.0000

. estat hettest, rhs

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variables: All independent variables

H0: Constant variance

   chi2(20) = 512.13
Prob > chi2 = 0.0000
```

Figure 11. Test for Heteroskedasticity

5. Conclusion

In conclusion, our regression model suggests that oil prices have a significant effect on sales. However, the model also indicates the presence of heteroscedasticity, which means that the variance of the error terms is not constant across observations. This may indicate that there are other factors at play that are not accounted for in the model.

To improve the accuracy and reliability of our model, we need to consider additional factors such as NPS and GDP. NPS can give us insights into the customer satisfaction level and loyalty, which can have a significant impact on sales. GDP is also an essential factor that can affect consumer behavior and purchasing power. By including these variables in our model, we can better understand the relationship between these factors and sales, leading to more accurate predictions and better decision-making.

It is worth noting that our OLS model showed some violations of the OLS assumptions, such as heteroscedasticity and potential multicollinearity. To address these issues, we could explore alternative modeling techniques such as multinomial classification and elastic net regression, which are more robust to these assumptions. Multinomial classification can help us deal with categorical dependent variables, while elastic net regression can handle multicollinearity and variable selection simultaneously.

Therefore, incorporating additional factors into our model and exploring alternative modeling techniques can help us obtain more accurate and reliable predictions.