

Forecasting of Liquefied Petroleum Gas (LPG) Refilling Plant Sales in Time Series Using Statistical Approaches and Machine Learning Techniques

Mary Jane C. Samonte and Jean Shermin B. Geronimo

School of Information Technology

Mapua University, Manila, Philippines

mjcsamonte@yahoo.com, jean8031978@yahoo.com

Abstract

The study focused on the conventional yet comprehensive use of forecasting in sales prediction. Businesses use forecasting to ascertain the allocation of their budgets or schedule for foreseen expenditures for a future time frame. In this study, forecasting is placed in the context of SG Trading which is a Filipino-owned company operating in Quezon City, Metro Manila. Methodically, the study makes use of both literature and actual experiment. The review was employed to check the prevailing existing significant literature to summarize the results. Similarly, the experiment was considered because it is the most satisfactory research method when exploring quantitative data (Brownlee, 2020). Its main goal was to gauge the performance of the classical statistic forecasting model, namely Auto-Regressive Integrated Moving Average, and Autoregressive and moving average, machine learning models, Support Vector Machine Regressor, Random Forest Regressor, and Gradient Boosting Regressor on the sales data obtained from the sales daily report of SG Trading. The results from the experiment revealed that XGBoost is the best suitable algorithm to forecast the LPG daily sales of SG Trading with the least error. In view of such findings, it is recommended to adopt and utilize the said forecasting model so as to have an effective planning of the LPG supply. Part of the limitations was the lack of variables; hence, the researchers have also suggested enhancing the sales data. In a nutshell, it is strongly enjoined to utilize the findings produced by various forecasting models for the improvement and sustainability of the business.

Keywords

Sales Forecasting, Forecasting Models, LPG Refilling, Time Series Models and Extreme Gradient Boosting.

1. Introduction

LPG is an essential source of modern, clean, and portable energy. It greatly helped improve people's lives who used to depend on wood and charcoal as their main source of domestic fuel. LPG is a remarkable energy source that plays a big role and brought social, economic, and environmental changes worldwide. The global LPG market is anticipated to grow extensively due to the expansive adoption of LPG as cooking fuel by residents. Furthermore, governments mostly in the main consuming economies of Southeast Asia, such as China, India, Indonesia, and the Philippines favors the initiative to endorse LPG as the alternative cooking fuel and auto fuel to replace traditional such as coal, wood, gasoline, kerosene, and diesel which foresee to push the demand for LPG. Due to insufficient natural gas infrastructure, high energy prices, issues with energy security, and strong economic growth show a great opportunity for LPG in the next 10 to 20 years in the Philippines (World LPG Association, 2015). The increasing demand and surge in the consumption of LPG and the profitability attached to it are the main drivers that gave businesses the urge to invest in the distribution of LPG in the country and resulted in the development in the competition of the LPG market. As LPG markets become increasingly global and with tighter levels of competition, it is most important to optimize the operational efficiency of organizations such as in the business of Liquefied Petroleum Gas (LPG) refilling. The LPG industry must extend its resources openly and consumers must be given a wide variety of choices, every market advantage an industry can thought-out will matter. The essentials in the decision making of management action is forecasting of sales, having as vital target to predict the factors that influence their operations and trade.

1.1 Objectives

This paper aimed to predict the LPG refilling sales in order to meet the following specific objectives:

- a. Determine the current methods for forecasting sales in LPG refilling or similar cases.

- b. Examine to establish the most acceptable performance measure of forecasting models.
- c. Identify the most applicable and eligible model for SG Trading.
- d. Understand the strategy in cleaning the data suitable for forecasting
- e. Perform best and most suitable model.
- f. Evaluate the performance of the selected forecasting model by comparing the success rate of prediction measure.

2. Literature Review

In the latest review on the analysis of methods and techniques for the prediction of natural gas consumption, the authors analyzed the papers by measures such as methods used to make predictions, independent variables used for modeling, the prediction area, and the prediction horizon. The majority of forecasting tools used in forecasting are closely linked to statistical approaches and artificial intelligence. In view of the aforementioned areas, it was found that the time series, regression analysis, fuzzy logic, artificial neural networks (ANN), Genetic Algorithms (GA), and expert systems are most significant. Furthermore, the econometric approach and so-called “end-use” approach are the ones most existence for medium-term and long-term forecasting.

Whereas, neural networks, various time series and regression models, statistical learning approaches, the so-called “similar day approach”, fuzzy logic models, and expert systems appear to be important for short-term forecasting (Šebalj, 2019). The use of classic forecasting tools combined with optimization is suggested when prediction is on national level Time series forecasting has always been the foundation in studying the behavior of any activity during a particular time interval. The movement of demand from the past and by taking into consideration other known circumstances in the future generate forecasts. Several machine learning models, such as, Auto-Regressive Integrated Moving Averag (ARIMA), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), Hy-brid Models and STL Decomposition (using ARIMA, Snaive, XGBoost) can be used to forecast sales, predict targets, and improve the strategy to enhance productivity in the near future of organizations and these methods have given the best forecasting accuracy (Udmale, 2017).

The methods and models for forecasting demand and consumption are variously identical because of the evaluation criteria. Most researchers agree that the foremost important measure of performance is the prediction accuracy of a model. However, for a given problem there is an applicable measure of accuracy that isn't considered to be suitable across all cases. An accuracy measure is described as a difference between the specified and the estimated value. There are a large number of performance measures, each with its advantages and limitations. The foremost common include: The mean absolute percentage error (MAPE), the mean squared error (MSE), the foundation mean squared error (RMSE), the sum of squared error (SSE), The mean absolute deviation (MAD), and also the coefficient of multiple determination (R^2). For forecasting sales using machine learning algorithms, the foremost common in measuring and evaluating performance is RSME, Mean Absolute Error (MAE), and R^2 . When comparing forecasts of LPG demand on one series, several common methods can be used: MAD, RMSE, MAPE, and therefore the mean error (ME) (Thomasson, 2017).

3. Methods

The results of the review were used as an input to the analysis and assessment of the research method experiment. The main goal of this experiment was to gauge the performance of the classical statistic forecasting model, namely Auto Regressive Integrated Moving Average (ARIMA) and Combined Autoregressive and moving average (ARMA), machine learning models, Support Vector Machine Regressor, Random Forest Regressor, and Gradient Boosting Regressor on the sales data obtained from the sales daily report of SG Trading. The results from the experiment were analyzed and compared to pick the most effective algorithm. The study was conducted in SG Trading located in Quezon City, Metro Manila. The target population of the study was 691 existing and walk-in customers which are dealers, retailers, commercial, household individuals, and other concerning consumers were included as study population.

4. Data Collection

4.1 Data Gathering and Analysis

Data required for Sales prediction is secondary quantitative variables. The data was collected from the submitted summary daily sales reports by the sales unit of SG Trading. The dataset employed in this study consists of the daily LPG sales from 2016 to 2020, which comprised of 104,350 transactions for 1,738 days. From these data, out of the 35 variables collected for analysis, only 15 variables passed the feature selection and were used in this study. The independent variables include the date, average unit price, mode of payment, payment term, commission, while the dependent variable is the total quantity of sales. External and exogenous variables like temperature, humidity, rainfall, working weekday, weekends, and holidays were also included in this experiment since these variables influence demand.

4.2 Data Preparation and Feature Selection

Data diagnostic was performed on each variable to determine any data issues. The incorrect values were validated by identifying the minimum and maximum values of the data. After determining the values below the minimum and above the maximum, the values were confirmed with the company if acceptable or not. Acceptable values were retained, and unacceptable values were removed. The researcher used filtering to check inconsistent data format and misspelled information. The incorrect data was corrected and transformed into its proper form and value. The researcher detected duplicate records using pivot and filtering. After spotting, the duplicate records were confirmed with the company. Confirmed duplicate records were removed. Missing values in the dataset were checked by the researcher using the summary function. All missing values were replaced with zero. Generally, two weight classifications were recorded within the dataset. The weights were combined into one and were used for the experiment. The dates for no sale for a selected category have been replaced with zero. Moreover, the sales parameter was used as a feature within the models. Data has been resampled from per record to daily and transformed into a supervised learning dataset.

Before modeling, the time series were tested and found to be stationary with trends and seasonality. The Augmented Dickey-Fuller test was performed to investigate stationarity in this experiment. The data was converted into stationary series using differencing twice.

4.3 Data Modeling

Statistical and Machine learning supervised approaches using labeled datasets were applied in the study. The forecasts were generated Sales from 2016 to 2020. Data were split 70% train and 30% test data. This split meant 1,264 days of data or 65,523 records as train and 541 days of data or 28,085 records to test.

A programming language for statistical computing and graphics was used for the modeling techniques. An application interface with an integrated development environment was used to simultaneously train several models and to run twenty-four (24) machine learning techniques. Time-series regression with Auto-Regressive Integrated Moving Averages, Autoregressive Moving Average, Support Vector Machine Regressor, Random Forest Regressor, and Gradient Boosting are performed on the dataset. Performances of the algorithms have been experimented, and the results are compared for selecting the best-performed algorithm for this dataset.

These quantitative forecasting methods engage expansive methods, and each method possesses properties, accuracies, and costs that ought to be accounted for when choosing a selected method within the specific domain for a specific objective. Quantitative methods also include, among others, multivariate analysis, decomposition methods, exponential smoothing, and therefore, the Box-Jenkins methodology. Statistic data collected at regular duration and interval, or partial data is required for many quantitative prediction problems, of which are gathered at one period of your time (Hyndman, 2018).

5. Results and Discussion

5.1 Results

The five-year monthly dataset was consolidated into one dataset as shown in Figure 1. Another aggregation made during the building of models was the summing up of all sales by date which is necessary for time series regression. Other features were also summed up for the exogenous averaging. The 70 percent initialization set was used to formulate the appropriate models for forecasting while the 30 percent second set called the test set was used to check the validity of the chosen models.

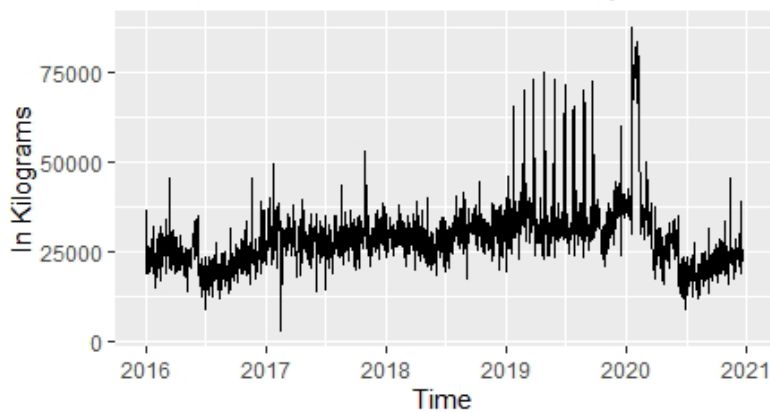


Figure 1: The time series plot of the LPG Daily Sales in SG Trading from 2016 to 2020

From 2016, there was an increase in the sales pattern in the months of January to October 2016 followed by a sudden decrease that rose after the first month in 2017, the third month experienced almost a constant stability and this gave way to a rising pattern that continued until the ninth month of 2019 and fell seriously and kept an undertone rise at that level till about the second month in 2020 and a fluctuating trend in the sales for LPG till about the end of 2020 as shown in Figure 2.

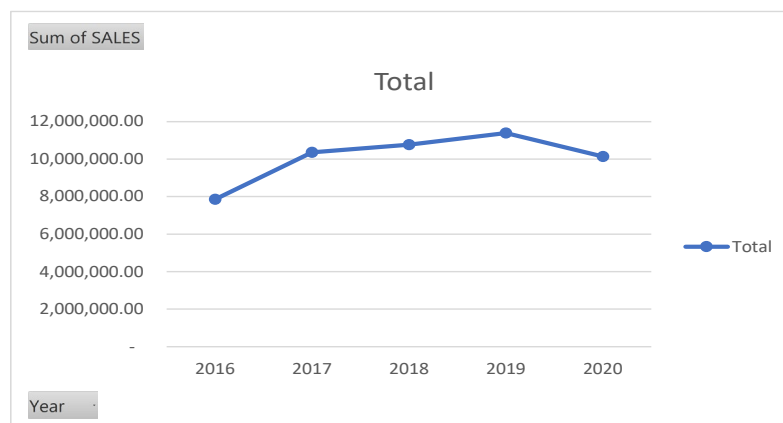


Figure 2: Yearly LPG Sales in SG Trading from 2016 to 2020

Total LPG sales consistently grew by 3.53 million kilograms or 6.98 percent from 7.86 million kilograms for the year 2016 to 11.38 million kilograms for the year 2019. In the year 2020, a sharp decrease of 1.24 million kilograms or 2.46 percent was recorded. The decrease for the year 2020 may be attributed to the strict quarantine implemented by the country brought about the global pandemic.

From 2016 to 2019, the month of December has the highest total LPG sales, which can be attributed to the Christmas season as shown in Figure 3, which is considered a cold month (U.S. Energy Information Administration, 2021). The months of July to October, which corresponds to the wet season in the Philippines and have cooler temperatures than other months (PAGASA, 2022), have the second-highest total LPG sales. In 2020, it was the month of February, which had the highest total LPG sales. This significant increase in sales may be attributed to the preparation of

distributors for the upcoming strict quarantine that was implemented in March 2020. Despite the rise in February 2020, the trend remained the same for the subsequent months.

For the year 2016 to 2020, the days of the week that has high total LPG sales were consistently observed during Sundays and mixed of Tuesdays, Wednesdays, and Thursdays as displayed in figure 4.

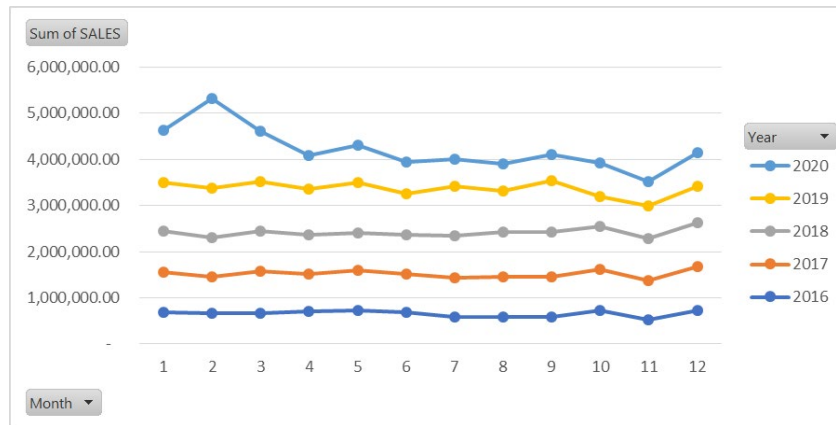


Figure 3: Monthly LPG Sales in SG Trading (2016 to 2020)

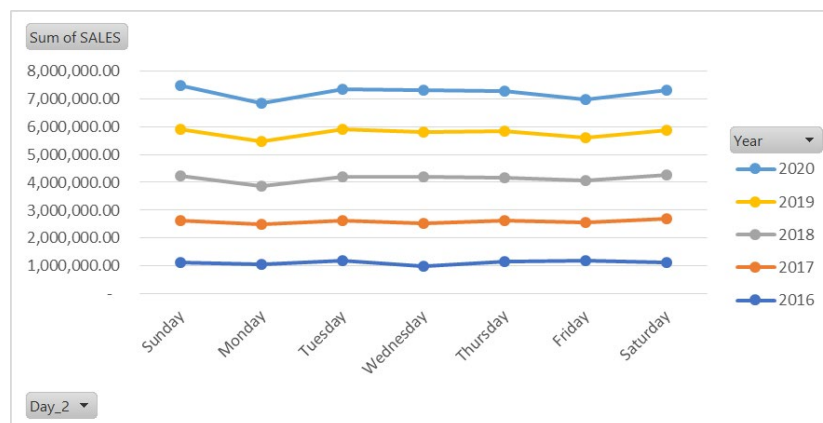


Figure 4: Days of Week LPG Sales Per Year in SG Trading (2016-2020)

Based on the statistics shown in Figure 5, the LPG sales of SG Trading decrease when the temperature is low, which is between the range of 24°C to 26 °C and follows the same trend when the temperature is high, usually ranging from 30°C to 33 °C. The LPG sales are highest when the temperature range is between 27°C to 29 °C, considered average temperature in the Philippines. Based from data shown in Figure 6, it is consistent with the study that temperature affects gas consumption (Botzen, 2020), and the LPG consumption declines with temperature. While the study reveals that LPG consumption is negatively affected by heat index fluctuations (Dacuycuy, 2020).

Each function shows the correlation between the LPG Daily sales values, a pattern of decreasing and increasing spikes showing both trend and seasonality in the LPG Daily sales in SG Trading. The autocorrelations show a way of decreasing and increasing values though not up to the initial one. The autocorrelation seems to be consistent above zero. The partial autocorrelation shows a pattern of decreasing values. However, there are increases in the various decreasing values, though not up to the initial ones, and consistent decreases towards zero.

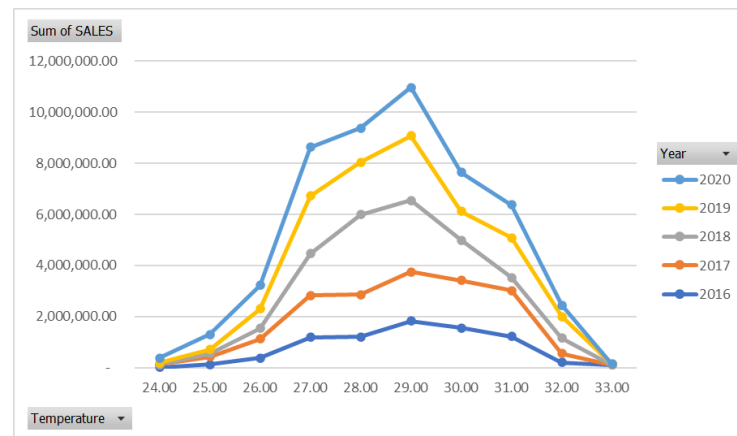


Figure 5: LPG Sales vs. Temperature (2016-2020)

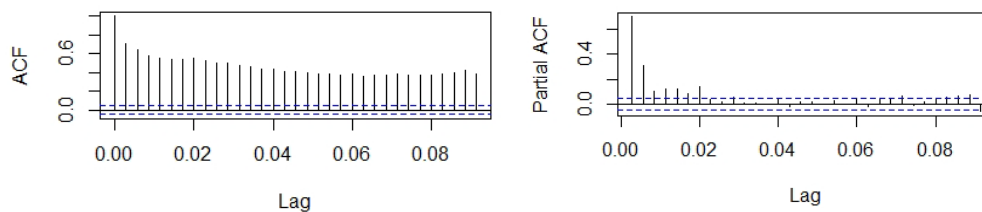


Figure 6: ACF and PACF plot of the LPG Daily Sales

5.1.1 ARIMA

Parameter estimates determined the coefficients of the time series equation that is generated from the data and the diagnostics test is used to check the correlation and significance of the residuals. The selected ARIMA model is denoted by ARIMA (2,2,2), the meaning of the parameters is $p=2$ is the order of autoregressive part, $d=2$ is the degree of differencing, and $q=2$ is the order of moving average part or represents the number of preceding/lagged values for the error term. Since ARIMA uses maximum likelihood for estimation, the coefficients are asymptotically normal. The number of times that the scores were differenced to make the process stationary, determines the value of d . Since the series was differenced twice, $d = 2$, and both linear and quadratic trends were removed. For nonstationary series, d values of 1 or 2 are usually adequate to make the mean stationary. The p -values for each coefficient were less than the significance level of 0.05 which means that the null hypothesis need not be rejected and the term in the model is statistically significant.

5.1.2 ARMA

The residuals of the ARMA (1,1) model seem to better follow a normal distribution compared to ARIMA (2,2,2). Likewise, the residuals of the ARIMA (1,1) model are less autocorrelated. The Mean Absolute Error (MAE) acquired

by ARIMA (1,1) on a time series regression test is 4633.45 and the Root Mean Squared Error (RMSE) acquired by ARMA (1,1) is 7253.08.

5.1.3 Random Forest Regressor

The Mean Absolute Error (MAE) acquired by a random forest model with mtry=10 and trees=500 on a 5-fold time series cross-validation test is 2408.66 and the Root Mean Squared Error (RMSE) acquired is 3259.34. Moreover, the R^2 for the training data acquired was 87.59 and 88.81 for the test data.

5.1.4 Support Vector Regressor

The Mean Absolute Error (MAE) acquired by the SVM model with type eps-regression and kernel=radial and trees=500 on a 50-fold time series cross-validation test is 2458.91. The Root Mean Squared Error (RMSE) acquired is 3893.82. The R^2 for the training data acquired was 85.28. and 83.66 for the test data.

5.1.5 Extreme Gradient Boosting Regressor

The Mean Absolute Error (MAE) acquired by the XGBoost model with parameter=booster on regression tests is 2119.68. Additionally, the Root Mean Squared Error (RMSE) acquired is 2975.20, while the R^2 for the training data acquired was 89.90 for the test data. The data modelling results show that importance of each variable in the selected model. Out of the ten variables used in the model, it was found that the unit price given to customers who purchased LPG for the refill of cylinders more than 3 kilograms capacity is the most significant which it comes to the prediction of sales.

The XGBoost Regressor has performed significantly well compared to the other models. This may be attributed to the regularization approach where the variance is reduced at the cost of some bias initiation which makes it robust to outliers and overfitting. Support Vector Regression has performed surprisingly well compared to the random forest, ARMA, and ARIMA, which may be linked to generalization capability as shown in Table 1. Kernel functions for the chosen parameters were set to linear, polynomial, radial, and sigmoid which may have produced better results. Random Forest Regressor has not shown a good performance, this may be because of the overfitting problem and may have given more preference to the hyperparameter which may be preventing from generalizing the model. ARIMA and ARMA produced bad results compared to others because of the overfitting problem, thus making the two harder to tune compared to Random Forest.

Table 1: Comparison of performance evaluation results

ALGORITHM	MAE	RSME
ARIMA(2,2,2)	6330.67	9579.85
ARMA(1,1)	4633.45	7253.08
RANDOM FOREST REGRESSOR	2408.66	3259.34
SUPPORT VERCTOR REGRESSOR (SVR)	2458.91	3893.82
EXTREME GRADIENT BOOSTED REGRESSOR (XGBOOST)	2119.68	2975.20

5.2 Discussion

Based on the literature review, the best current methods in forecasting sales in LPG refilling or similar cases are the statistical approach and machine learning algorithms. Existing literatures have revealed several strategies in cleaning the data to be suitable for forecasting. For time series data, it is important to check the trend, seasonality, and stationarity of the data by using visualization tests or statistical tests such as the Augmented Dickey-Fuller Test. Autocorrelation and partial autocorrelation should also be checked using the ACF plot and PACF plot.

XGBoost is the best suitable algorithm to forecast the LPG daily of SG Trading. In this experiment, XGBoost has least error in forecasting the sales when compared to the Support Vector Machine, Random Forest, ARMA, and ARIMA. This is because of the regularization, where slack variables are added to avoid over-fitting. Average RMSE across the 5-fold time series cross validation is 2975.20 which is quite outstanding and Average MAE is 2119.68. ARIMA has shown worst performance compared to the other algorithms, RMSE across the 5-fold time series cross-

validation is 9579.85 which is worst, and MAE is 6330.67. The performances of all the models have been discussed in the section where the model selection for the LPG daily sales data was made.

This sales data analysis gave information on how much revenue an average customer generates during their lifetime with the company. This was used to predict future revenue. With the customer lifetime value metric, the company can make informed decisions on how much can be spent on acquiring new customers. The sales data also revealed that the number of customers decreased during the last two years. This information was used to prepare customer strategic objectives and company goals in terms of customer retention, increase partner agents, and third-party vendors. The forecast results further show that the company will have a reoccurring revenue and this information was used to plan for the internal/operational strategic objectives like innovation, product development, an improvement on infrastructure, and expansion.

6. Conclusion

Data analysis and forecasting support innovation, foster more appropriate and real-time decisions, and would subsequently lead to the improvement in organizational performance. Sales forecasting is both a vital and pivotal part of the financial planning of business for any company. It serves as a self-assessment tool that uses statistics of the past and the current sales to predict future performance. Sales forecasting takes part in optimizing the LPG sales process. Financial and sales planning along with the sales forecasts give the information needed to project the revenue as well as the profit. After performing the various statistical tests and performance metrics, it was found that Extreme Gradient Boosting Regression is a suitable algorithm in conformity with the chosen dataset and thus accomplishing the aim of this study. The recommended model for predicting future values for the LPG sales of SG Trading is Extreme Gradient Boosting.

References

- Brownlee, Jason., *11 Classical Time Series Forecasting Methods in Python (Cheat Sheet)*.
<https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/>, 2020.
- Botzen, W.J.W.; Nees, T.; Estrada, F., *Temperature Effects on Electricity and Gas Consumption: Empirical Evidence from Mexico and Projections under Future Climate Conditions*. Sustainability 2021, 13, 305, 2020.
<https://doi.org/10.3390/su13010305>.
- Chakure, A., *Random Forest Regression. Star it up: The Startup*. <https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f>, 2019.
- Dacuycuy, C.B., *Energy Consumption, Weather Variability, and Gender in the Philippines: A Discrete/Continuous Approach*. The Research Information Staff, Philippine Institute for Development Studies.
<https://pidswebs.pids.gov.ph/CDN/PUBLICATIONS/pidsdps1706.pdf>, 2017.
- Great Learning Team., *Support Vector Regression in Machine Learning*. Great Learning.
<https://www.mygreatlearning.com/blog/support-vector-regression/>, 2020.
- Hyndman, R., Koehler, A.B., Ord, J.K., Snyder, R.D., *Forecasting with exponential smoothing: the state space approach*. Springer, Berlin, 2008.
- Hyndman, R.J., & Athanasopoulos, G., *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2, 2018.
- Influxdata, *Time series analysis methods*. <https://www.influxdata.com/time-series-analysis-methods/>. (visited on 04/13/2021), 2021.
- PAGASA, *Climate of the Philippines*. Url: <https://kidlat.pagasa.dost.gov.ph/index.php/climate-of-the-philippines>, 2022.
- Šebalj, Dario & Mesarić, Josip & Dujak, Davor., *Analysis of Methods and Techniques for Prediction of Natural Gas Consumption: A Literature Review*. Journal of Information and Organizational Sciences. 43. 99-117. 10.31341/jios.43.1.6, 2019.
- Sarkar, P., *What is Logistic Regression in Machine Learning*. Knowledge Hut.
<https://www.knowledgehut.com/blog/data-science/logistic-regression-for-machine-learning>, 2019.
- Thomasson, S.A., *Improving forecast performance of LPG demand*. University of Twente.
<http://purl.utwente.nl/essays/73069>, 2017.
- Udmale, S. and Sambhe, V., *Forecasting of sales by using fusion of Machine Learning techniques*. International Conference on Data Management, Analytics and Innovation (ICDMAI) Zeal Education Society, Pune, India, 2017.

U.S. Energy Information Administration, *Independent Statistics and Analysis: Natural gas explained Factors affecting natural gas prices*. Washington, DC. <https://www.eia.gov/energyexplained/natural-gas/factors-affecting-natural-gas-prices.php>, 2021.

World LPG Association, *LPG-and-the-Global-Energy-Transition*. <https://www.wlpga.org/wp-content/uploads/2015/05/LPG-and-the-Global-Energy-Transition.pdf>, 2015.

Biographies

Mary Jane C. Samonte has a double bachelor's degree in computer education and information technology. She also has two post graduate degree; Information Technology and Computer Science. She finished her Doctor in IT with a study focusing in Deep Learning. She has a wide range of research interests that are centered around educational technologies, gamification, mobile and ubiquitous learning, digital game-based learning, artificial intelligence in education, e-health, assistive technology, natural language processing, green computing and data analytics-based studies.

Jean Shermin B. Geronimo graduated from De La Salle University, Manila, Philippines with B.S. degree in Mathematics major in Actuarial Science with Specialization in Statistics and earned her Masters in Information Technology in AMA University, Metro Manila, Philippines. She is a member of the Philippine Computer Society (PCS) and also an active member in the Analytics Association of the Philippines (AAP). Her research interest includes Data Analytics and Forecasting.