

A Framework for Machine Learning Based Demand Forecasting in LPG Distribution

Astitva Keshari¹, Arya Verma², Riddhi Jagdale³, Abhinav Raj⁴

Student, Department of Information Technology

Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India^{1,2,3,4}

Abstract: *Liquid Petroleum Gas (LPG) is the primary cooking fuel used by many households in India, and the local distributor is the last link in the chain between the supplier and the consumer. The majority of the distributors still make their indents, their godown stocks, and the labour force on delivery based on their experience and simple averages which is of no avail during the peak of winter, the festival period and during disturbance of supply. This paper summarizes the various techniques available for demand forecasting including statistical models, machine learning and deep learning, applying them specifically to the studies of the gas and LPG demand. The major gaps at the distributor level are highlighted, such as the absence of data at the distributor level, the absence of Indian calendar features, and the weak linkage between forecasts and stock decisions. A five-layered framework is then introduced that spans data acquisition, feature engineering, forecasting models, evaluation and decision support. Designed for a small distributorship having godown, showroom and delivery staff, the framework translates the forecasts straight into safety stocks, reorder points and stock allocation.*

Keywords: LPG distribution, demand forecasting, machine learning, time series, inventory management, decision support

I. INTRODUCTION

The LPG consumption data for the nation is shared by the Petroleum Planning and Analysis Cell [2] and the LPG consumption in India has been increasing rapidly in the past decade, partly due to schemes like the Pradhan Mantri Ujjwala Yojana (PMUY) [1]. The figures are preceded by thousands of local distributors who are charged with filling the cylinders at the oil marketing company (OMC) plants, storing them and bringing them to the customers. A distributor has to make decisions on the number of cylinders to indent each week, on splitting stock between godown and showroom, and on employing the number of delivery staff, every week. When demand is underestimated, it is a source of booking backlogs and complaints, while the opposite, over-estimating, leads to the capital tying up in the filled cylinders and increases the storage risk. The sales data is typically used for the past month, and decisions are made based on experience. These are sometimes taken to be a reflection of that month's sales. These are generally taken to be representative of sales from the previous month; and experience is used to make decisions. The simple averages cannot be combined to reflect demand because it also responds to weather, festivals, and price changes and new connections. While machine learning (ML) has proven to be a good fit for retail, electricity and natural gas forecasting, there is limited work on the level of an LPG distributor. This paper contributes in three ways: (i) it provides focused overview of forecasting approaches for LPG demand, (ii) points out a number of research gaps related to distributor-level forecasting and (iii) presents a five-layer framework for connecting forecasts with stock and staffing decisions.

II. LPG DISTRIBUTION AT THE DISTRIBUTOR LEVEL

Bulk LPG moves from refineries and import terminals to OMC bottling plants, where it is filled into 14.2 kg domestic, 5 kg and 19 kg commercial cylinders. Distributors receive filled cylinders by truck and return empties for refilling, as



shown in Fig. 1. A typical distributor keeps stock at two points: a licensed godown and a customer-facing showroom, each with its own register. Customers book refills through IVRS, SMS, apps or WhatsApp, and delivery staff (hawkers) deliver the cylinder, collect the empty and the payment, and confirm delivery with an OTP-based code.

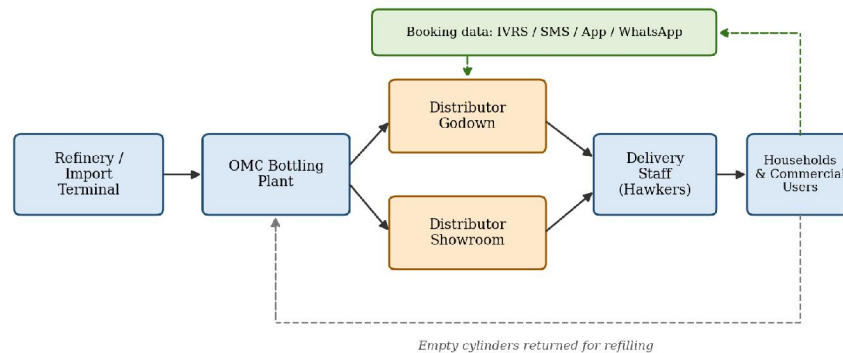


Fig. 1 Simplified LPG cylinder supply chain with booking data flow and empty cylinder return loop
Demand at this level rises in winter, spikes around festivals and weddings, grows slowly with new connections and shifts when OMCs revise prices. Supply problems at the plant can also create backlogs that appear later as sudden delivery peaks.

III. LITERATURE REVIEW

A. Statistical Methods

Statistical Methods Other commonly used baselines are ARIMA and seasonal ARIMA (sARIMA) models [3] which rely on past values and errors of a series. Exponential smoothing methods, like Holt-Winters, do not require much data, and they are very flexible in the sense that they adapt rapidly, but are not able to consider external variables. A rule of thumb suggested by Hyndman and Athanasopoulos [4] is to always compare the new models with some simple models through time series cross-validation and scaled error metrics like MASE. For the case of Indian festivals date varying with a year, explicit modelling of trend, seasonality and holidays was proposed by Taylor and Letham [5] for which Prophet was developed.

B. Machine Learning Methods

Random forest [6] and gradient boosting libraries like XGBoost [7] and LightGBM [8] approach forecasting as a supervised learning problem using the lag and calendar features. They have a non-linear relationship, work with small data sets, and LightGBM was the winning algorithm in the M5 retail forecasting competition [9]. Their sole disadvantage is that tree models cannot predict an increasing trend. When applying the ML methods to supply chain demand, Carbonneau et al. [10] demonstrated that the gains achieved were relatively modest when compared to traditional methods and it is important to carefully compare the two.

C. Deep Learning and Hybrid Methods

LSTM networks [11] model long-term patterns in a sequence of data and DeepAR [12] can make probabilistic forecasts for several related sequences, appropriate for inventory planning. Known inputs into the future like holidays can be used as inputs to the Temporal Fusion Transformer [13]. The models require more data and tuning as compared to tree models. The M4 competition [15] demonstrated combinations and hybrids tend to be the most accurate, while Zhang [14] indicated that a combination of ARIMA and a neural network on the ARIMA residuals is likely to be more accurate.



D. Gas and LPG Demand Forecasting

Neural networks and regression were the most common models used for natural gas, where temperature was the most common input, but most of the studies were on a city or national scale, as mentioned by Šebalj et al. [16]. Little research exists on LPGs. Panichatsadorn and Ketcham [17] applied time series and neural networks to an LPG station. Not sufficient was the only information available, the sales history in Thailand. Correia et al. [18] predicted LPG cylinder sales and empty returns for an LPG company in Portugal using selected models based on ensemble modeling techniques, such as statistical, regression and neural models. Yoshida et al. [19] predicted gas consumption in households in Japan using SVR, random forest and k-nearest neighbour models, and estimated the lengths of cylinder replacement routes based on predictions of gas use, and over a field test, reduced the length of these routes. By providing explanations of model outputs, SHAP [20] can enhance trust in small businesses. The key works are summarised in Table I.

TABLE I Summary of Key Reviewed Works

Ref.	Method	Domain / Data	Strength	Limitation
[3]	ARIMA / SARIMA	General time series	Standard baseline	Linear, manual tuning
[5]	Prophet	Business time series	Holiday effects	Misses sharp spikes
[7], [8]	XGBoost, LightGBM	Tabular, retail sales	Accurate, uses features	No trend extrapolation
[11]	LSTM	Sequential data	Long-term patterns	Needs more data
[12]	DeepAR	Many related series	Probabilistic output	Needs many series
[14]	ARIMA + ANN	Benchmark data	Linear + non-linear	Two-stage tuning
[17]	Time series + ANN	LPG station, Thailand	LPG specific	Single site
[18]	Ensemble + Monte Carlo	LPG cylinders, Portugal	Sales and returns	Monthly, company level
[19]	SVR, RF, kNN + routing	LPG customers, Japan	Forecast linked to delivery	Needs smart meters

IV. RESEARCH GAPS

- Almost no published work forecasts refill demand for a single LPG cylinder distributor, and no public dataset exists.
- Indian factors such as moving festival dates, wedding seasons, price revisions and PMUY growth are rarely modelled together.
- Most studies stop at error values; few link forecasts to stock decisions, and none handle two stock points such as a godown and a showroom.
- Stock-outs hide real demand (censoring), and advanced deep models are too heavy for small distributors.

V. PROPOSED FRAMEWORK

The proposed framework has five layers, shown in Fig. 2. It is designed for a single distributorship and can be built with spreadsheets for data entry and Python libraries for modelling.



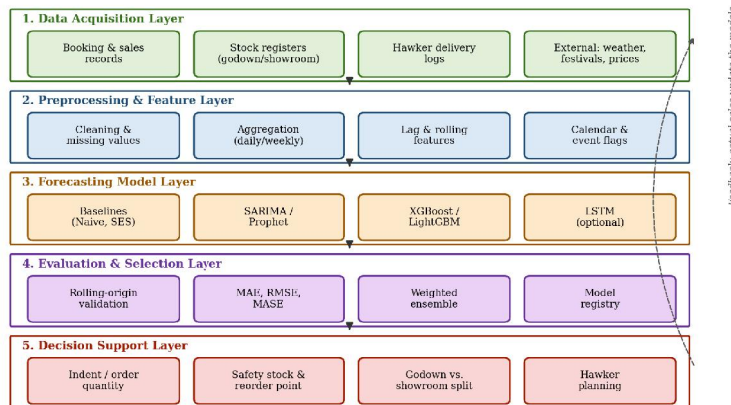


Fig. 2 Proposed five-layer framework for ML-based demand forecasting in LPG distribution

A. Data Acquisition

Internal data comprises of daily bookings and deliveries by cylinder type, stock received, godown stock and showroom stock, empties returned, new connections and hawker delivery logs. External data comprises the temperature data and calendar of festivals and dates of revising OMC prices. Each of all the sources is combined on a single table every day, based on the cylinder type.

B. Preprocessing and Features

Days missing and entries that are off are cleaned, but real spikes are retained and identified. Bookings are recorded, even during stock-outs, thus the bookings are used as the demand signal and the number of bookings, not deliveries, reduces censoring. The features presented are shown in Table II.

TABLE II Proposed Feature Set

Feature Group	Examples	Source
Lag and rolling	Bookings 1, 7, 28 days ago; 7-day mean	Booking records
Calendar and events	Day of week, month, days to festival, holiday flag	Local calendar
Weather	Daily minimum and average temperature	Weather service
Customer base	Active and new connections	Connection register
Price and supply	Days since price revision, stock-out flag, backlog	OMC notices, stock registers

C. Forecasting Models

The models are developed in steps: a baseline model using seasonal naïve and exponential smoothing, then SARIMA and Prophet including Indian festival holidays as [5] or LightGBM or XGBoost on the Table II features [7], [8]. The target is based on bookings per 1,000 connections, to account for trends. If sufficient data is available, then only LSTM or DeepAR [11] [12] is used. The final forecast is a combination of the best models, weighted by their performance over the last 1–4 weeks.

D. Evaluation

Rolling-origin validation is used instead of random splits. Accuracy is measured using MAE, RMSE and MASE, where y_t is actual demand, $\hat{y}_t$ is the forecast and m is the season length. A MASE below one means the model beats the seasonal naïve forecast [4].

$$MAE = (1/n) \sum |y_t - \hat{y}_t|, \quad RMSE = \sqrt{(1/n) \sum (y_t - \hat{y}_t)^2}$$

$$MASE = MAE / [(1/(n-m)) \sum |y_t - y_{t-m}|]$$



E. Decision Support

Forecasts are converted into actions. Safety stock (SS) and reorder point (ROP) are computed as below, where $\hat{D}_L$ is forecast demand over the lead time, σ_L is the forecast error deviation over the lead time and z is the service factor (1.65 for about 95 percent service).

$$SS = z \times \sigma_L, \quad ROP = \hat{D}_L + SS$$

If the stock level falls below the ROP, an indent quantity is recommended and compared to the godown storage limit. The forecast is split between godown and showroom according to their usual sales share and the daily forecast divided by the average godown delivery per hawker results in the number of hawkers required. The weekly time frame is illustrated in Fig. 3. Each forecast is explained by SHAP values [20] and the owner retains the decision.

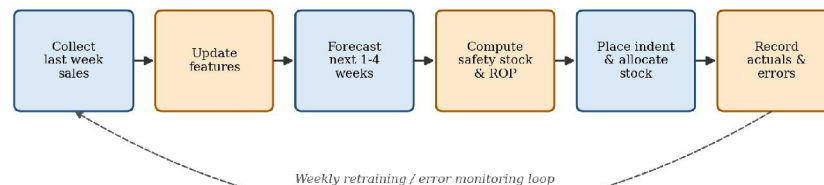


Fig. 3 Weekly operational workflow of the proposed framework

VI. DISCUSSION AND FUTURE WORK

The framework is supposed to minimise stock-outs and overstock, lessen the time lag between the godown and the showroom during peak hours and minimise the number of times the goods are being moved from the godown to the showroom. It has a staged design where a distributor can begin with booking history. Poor data in manual registers, short history, supply issues not in the distributor's control and policy changes that necessitate retraining are limitations. The structure is theoretical and is yet to be validated with real data. Future work involves implementation on actual data from distributors, creating an anonymised benchmark data set, aggregating data at distributor level for use in global models, and integration with hawker route planning as done in [19].

VII. CONCLUSION

This paper reviewed forecasting methods relevant to LPG demand and found that very little work targets the local distributor, where daily decisions on indents, stock location and staffing are made. A five-layer framework was proposed that combines practical data collection, Indian calendar features, staged models, time-aware evaluation and a decision layer producing safety stock, reorder points, stock allocation and hawker plans. Testing the framework on real distributor data is the main next step.

Acknowledgment

The authors thank the Department of Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, for its guidance and support during this work.

REFERENCES

1. Pradhan Mantri Ujjwala Yojana, Ministry of Petroleum and Natural Gas, Government of India. [Online]. Available: <https://www.pmu.gov.in>
2. Petroleum Planning and Analysis Cell (PPAC), Ministry of Petroleum and Natural Gas, Government of India. [Online]. Available: <https://ppac.gov.in>



3. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ, USA: Wiley, 2015.
4. R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: <https://otexts.com/fpp3/>
5. S. J. Taylor and B. Letham, "Forecasting at scale," The American Statistician, vol. 72, no. 1, pp. 37–45, 2018.
6. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
7. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
8. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems 30, 2017, pp. 3146–3154.
9. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," International Journal of Forecasting, vol. 38, no. 4, pp. 1346–1364, 2022.
10. R. Carbonneau, K. Laframboise, and R. Vahidov, "Application of machine learning techniques for supply chain demand forecasting," European Journal of Operational Research, vol. 184, no. 3, pp. 1140–1154, 2008.
11. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
12. D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," International Journal of Forecasting, vol. 36, no. 3, pp. 1181–1191, 2020.
13. B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748–1764, 2021.
14. G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, pp. 159–175, 2003.
15. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 competition: 100,000 time series and 61 forecasting methods," International Journal of Forecasting, vol. 36, no. 1, pp. 54–74, 2020.
16. M. Šebalj, J. Mesarić, and D. Dujak, "Predicting natural gas consumption – a literature review," in Proc. 28th Central European Conf. Information and Intelligent Systems (CECIIS), Varaždin, Croatia, 2017, pp. 293–300.
17. W. Panichatsadorn and M. Ketcham, "Demand forecasting for LPG station by time series and artificial neural network," Journal of Information Science and Technology, vol. 7, no. 1, pp. 42–49, 2017.
18. Correia, C. Lopes, E. Costa e Silva, M. Monteiro, and R. B. Lopes, "A multi-model methodology for forecasting sales and returns of liquefied petroleum gas cylinders," Neural Computing and Applications, Springer, 2020.
19. Yoshida et al., "Long-term optimal delivery planning for replacing the liquefied petroleum gas cylinder," arXiv preprint arXiv:2112.12530, 2022.
20. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

