Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/AF6.20250086
Published: April 8, 2025

Generative AI for Inventory Management: A VAE-Based Approach to Demand Modeling

Mohammad Arbabian, and Naga Vemprala
Assistant Professor
Pamplin School of Business
University of Portland
Portland, OR
USA
arbabian@up.edu, vemprala@up.edu

Naveen Gudigantala
Associate Professor
Pamplin School of Business
University of Portland
Portland, OR
USA
gudigant@up.edu

Abstract

Variational Autoencoders (VAEs) represent a powerful Generative AI tool that learns the underlying structure of input data and generates new data. This paper applies VAEs to model complex, real-world demand distributions in the context of inventory management. Existing research often relies on traditional methods, such as linear regression or Gaussian mixture models, which frequently struggle to capture the non-linear and multimodal nature of actual demand patterns. VAEs utilize deep neural networks to learn intricate patterns within data and map them to a regularized latent space, enabling the generation of novel demand scenarios. Consequently, using VAEs could result in finding the optimal order quantity more accurately. We hypothesize that VAEs empower businesses to make data-driven decisions to optimize inventory levels and minimize costs, providing a robust and adaptable framework for inventory management compared to traditional approaches. To test our hypothesis, we simulate ten sets of demand distributions with complex patterns and shapes and use VAEs to model these demand patterns. Based on the distributions identified by the VAEs, we calculate the optimal order quantity and expected profits. We then compare these results to scenarios where demand is assumed to follow traditional distributions, including Normal, Poisson, Erlang, and Uniform distributions. Our findings reveal that, in most cases, the VAE-based approach outperforms the approaches that use traditional distributional assumptions. This research demonstrates the potential of VAEs to contribute to inventory management by providing a more accurate and flexible approach to modeling demand variability, ultimately leading to significant cost savings and improved operational efficiency.

Keywords

Variational Autoencoders, Inventory Management, Demand Forecasting, Generative AI, and Probabilistic Modeling

1.Introduction

The increasing complexity and dynamism of business environments necessitate effective inventory management for firms to maximize profitability and ensure quality operations. However, firms face significant challenges in aligning inventory levels with uncertain demand, often resulting in overstocking or stockouts (Hao 2024). Traditional models, such as the Newsvendor model, remain widely used due to their analytical tractability and practical applicability. However, these models rely on the assumption of specific probability distributions to characterize demand. This reliance can be problematic as it often oversimplifies the demand structure and ignores the complex and inherent uncertainties present in real-world scenarios, including demand shocks, supply chain disruptions, and logistical bottlenecks (Saghafian and Brian 2016). A more flexible and data-driven approach to estimating demand distributions is necessary to improve decision-making in inventory management.

In recent years, advances in Generative AI have provided new opportunities to enhance traditional inventory models by learning demand patterns directly from data. Can advanced Generative AI capabilities address the limitations of conventional demand estimation methods? We propose a novel approach by augmenting the classical Newsvendor model with the advanced capabilities of Variational Autoencoders (VAEs), a deep generative model designed to capture and learn complex probability distributions from historical data. Unlike traditional approaches that assume demand follows a predefined parametric distribution (e.g., normal, Poisson), VAEs provide a data-driven method for estimating the underlying demand structure without restrictive assumptions. This approach enables the generation of synthetic demand scenarios that are more representative of real-world complexity, allowing businesses to develop more robust inventory policies that adapt to dynamic market conditions (Pinheiro Cinelli et al. 2021).

Our research objective is to evaluate the effectiveness of demand distributions generated using VAEs compared to traditional common normal distribution assumptions within the Newsvendor context. We employ a simulation-based comparative framework in which we synthesize demand datasets that mimic real-world variability, training a VAE to extract latent representations of demand patterns. By leveraging the probabilistic nature of VAEs, we generate multiple demand scenarios, each reflecting different plausible demand realizations. Using these generated scenarios, we determine the optimal order quantity by applying the critical ratio derived from the Newsvendor cost parameters. We then assess the expected profit outcomes under both VAE-based and traditional demand estimation approaches. Our study contributes to the field of inventory management by demonstrating the advantages of incorporating Generative AI methods in addressing demand uncertainty. We argue that this enhanced approach improves inventory decision-making, leading to better financial performance and reduced risk for firms.

Variational Autoencoders (VAEs) offer remarkable flexibility in capturing complex data distributions, making them particularly useful for modeling demand uncertainty in real-world settings. Traditional statistical methods, such as linear regression or Gaussian mixture models, often struggle to represent non-linear, multimodal, or non-Gaussian demand distributions effectively. These limitations arise because conventional models rely on fixed functional forms and assumptions that may not align with the true underlying demand patterns. In contrast, VAEs leverage deep neural networks as function approximators for both the encoder and decoder, enabling them to learn intricate relationships and dependencies within demand data.

The VAE framework consists of an encoder-decoder structure, where the encoder maps observed demand data into a latent space, capturing essential features of the distribution. This latent space is regularized to follow a simple prior distribution, typically Gaussian, ensuring smooth and efficient sampling. The decoder then reconstructs the observed data or generates new demand scenarios by sampling from this latent space (as shown in Figure 1). Unlike traditional demand estimation methods, VAEs can model non-linear and multimodal demand structures, allowing for the identification of distinct demand clusters and patterns. This flexibility is particularly valuable in inventory management, where demand variability is influenced by factors such as seasonality, economic fluctuations, and shifting consumer preferences.

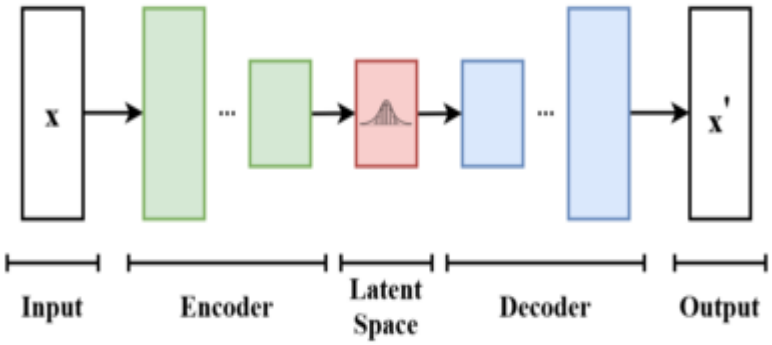

Figure 1. Visual representation of VAEs

One of the key advantages of VAEs is their ability to model latent variables probabilistically, thereby incorporating uncertainty directly into demand estimation. This feature is crucial in real-world inventory settings, where demand uncertainty is a persistent challenge. By learning a probabilistic representation of demand, VAEs enable businesses to generate diverse synthetic demand scenarios that reflect possible future outcomes more accurately than traditional distributional assumptions. Furthermore, the decoder network can generate plausible yet novel demand samples beyond the training data, improving generalizability and robustness in inventory decision-making.

In the context of the Newsvendor problem, VAEs offer several advantages over conventional demand modeling techniques. First, they eliminate the need for strong parametric assumptions about demand, instead learning distributions directly from data. Second, VAEs provide a structured approach to generating multiple demand scenarios, which can be used to derive more robust inventory policies. Third, their ability to capture complex demand distributions allows businesses to better assess the risk of stockouts or excess inventory, leading to improved profitability. By integrating VAEs into the Newsvendor framework, firms can make more informed and data-driven inventory decisions, ultimately reducing costs and enhancing service levels. Through the combination of deep learning and probabilistic inference, VAEs offer a powerful approach for demand estimation in inventory management. By leveraging this technology, businesses can transition from static, assumption-driven forecasting models to adaptive, data-driven frameworks that better reflect the complexities of real-world demand. This study highlights the potential of VAEs in enhancing inventory decision-making and provides a foundation for future research on integrating Generative AI with supply chain optimization.

2.Literature Review

Our study lies at the intersection of inventory management, artificial intelligence, and VAEs. Below, we review the literature related to each topic.

2.1 Inventory Management

Inventory management, with its implications for organizational efficiency and profitability, has been at the forefront of research in Operations Management and Supply Chain Management. To determine optimal order quantity under uncertain demand, the Newsvendor model is traditionally used (Qin et al. 2011). This model attempts to achieve a balance between stockouts and overstocking by using a critical ratio to recommend inventory decisions. To improve the model's decision-making under uncertainty, researchers have incorporated various demand distributions, such as Normal, Poisson, and Exponential distributions (Rossi et al. 2014). However, real-world demand data often includes a level of complexity and variability that may not be captured by traditional inventory models relying on simplifying assumptions. Qin et al. (2011), acknowledging the limitations of such assumptions, calls for methods that incorporate the non-linear aspects of real-world data. Subsequently, recent research is beginning to conceptualize new methods that fuse statistical models with machine learning approaches, transitioning towards developing more effective inventory management systems (Praveen et al. 2020).

2.2 Artificial Intelligence Applications in Inventory Management

The emergence of Artificial Intelligence and Machine Learning (AI/ML) has influenced many domains (Golchin and Riahi, 2021), and inventory management is no exception. Multiple research studies have begun to apply AI technologies to predict demand, optimize inventory and business processes, and contribute to improved decision-making in firms. For example, AI-based deep learning methods have outperformed traditional forecasting methods by better modeling the complex relationships within time-series data (Lim and Zohren 2021; Golchin and Rekabdar, 2024). AI-driven models ingest large amounts of data and allow for modeling complex relationships – steps typically missed by traditional methods. Mohammadi et al. (2024) applied a specific type of AI method, reinforcement learning, to solve a perishable inventory problem, and showed that their A2C algorithm is not only successful, but that it also outperforms the current policy. Therefore, the emergence of research in data-driven methodologies highlights the need for adaptive systems that can model complex patterns in demand or respond to unexpected supply chain disruptions.

Several studies have explored data-driven approaches to solving the Newsvendor problem by estimating demand distributions from historical data. Müller et al. (2017) propose using empirical distributions derived from observed demand data to optimize order quantities, while Simchi-Levi and Zhao (2021) introduce a machine learning-based approach that integrates demand data directly into the decision-making process. Araújo et al. (2022) provide a systematic literature review categorizing various data-driven Newsvendor models, emphasizing methods that do not assume predefined demand distributions. Feng et al. (2023) investigate machine learning techniques for demand estimation, proposing models that use historical data without requiring prior distributional knowledge. Additionally, Feng, Cai, and Chen (2024) leverage sentiment analysis on textual reviews to refine demand forecasts in the Newsvendor context. While these studies emphasize data-driven demand estimation, our work differs in that we employ Variational Autoencoders (VAEs) to learn complex demand distributions. Unlike traditional machine learning approaches, VAEs provide a generative framework capable of capturing multimodal and non-linear demand patterns, enhancing inventory decision-making in uncertain environments.

2.3 Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) offer remarkable flexibility in capturing complex data distributions, which is particularly advantageous when modeling phenomena like demand distributions in real-world settings. Traditional methods, such as linear regression or Gaussian mixture models, often struggle to represent non-linear, multimodal, or non-Gaussian distributions effectively. VAEs overcome these limitations by employing deep neural networks as function approximators for both the encoder and decoder. This architecture enables VAEs to learn intricate relationships and dependencies within the data, mapping them to a latent space that captures the essential features of the underlying distribution. The latent space is regularized to follow a simple prior distribution, typically Gaussian, which allows for efficient sampling and interpolation. By decoding these latent representations, VAEs reconstruct the observed data, capturing its variability and complexity in ways that traditional methods cannot achieve (Pinheiro Cinelli et al. 2021).

The flexibility of VAEs is further enhanced by their ability to model latent variables probabilistically, thereby incorporating uncertainty into their representations. This is particularly useful for real-world datasets, where demand patterns might exhibit high variability due to factors like seasonal fluctuations, economic trends, or customer preferences. The probabilistic nature of VAEs allows them to approximate distributions with multiple modes, representing distinct clusters or patterns in the data. Additionally, the decoder network can generalize beyond the training data, generating plausible yet novel samples that adhere to the learned distribution (Vázquez-García et al. 2025). This capability is especially helpful in the context of inventory management, where understanding the demand distribution is crucial for determining the optimal order quantity. By accurately modeling demand variability, VAEs provide insights into the probability of stockouts or overstocking, enabling businesses to make data-driven decisions that minimize costs and improve service levels. Through their combination of deep learning and probabilistic inference, VAEs offer a powerful and flexible framework for understanding, predicting, and utilizing complex demand distributions in inventory management.

VAEs have been shown to be valuable in domains beyond inventory management, such as anomaly detection, time-series analysis, and supply chain management (Khlie et al. 2024; Golchin & Rekabdar 2024). Their ability to improve demand forecasting stems from their flexibility in capturing complex relationships in real-world data. Combining VAEs with inventory management frameworks allows researchers to create models that account for both volatility in demand data and dynamic market conditions, thereby contributing to optimization of the inventory management process.

Thus, the literature points to a new direction of using Generative AI methods, such as VAEs, with classical inventory management frameworks to address potential limitations related to modeling demand uncertainty. These studies also show the potential of advanced machine learning methods to better manage inventory practices and drive positive firm-level outcomes.

1. Methods

The Newsvendor problem is a classic model in inventory management, used to determine the optimal order quantity for a single-period product under uncertain demand. The model balances the costs associated with understocking, which leads to lost sales or backorders, and overstocking, which incurs holding costs. Given

- c: Cost per unit ordered,
- p: Selling price per unit,
- b: Backorder penalty cost per unit,
- h: Holding cost per unit of leftover inventory,
- D: Demand, which is a random variable,
- *F(D):* CDF of the demand,

the objective is to maximize the expected profit by identifying the optimal order quantity (Q^*).

A critical factor in solving the Newsvendor problem is the accurate estimation of the demand distribution. The optimal order quantity is derived based on the critical ratio, which depends on the cumulative distribution function (CDF) of demand. If the demand distribution is misestimated, the decision-maker risks either underestimating demand, leading to missed revenue opportunities, or overestimating it, resulting in excessive inventory and higher holding costs. Therefore, precise knowledge of the demand distribution is fundamental to the profitability and efficiency of inventory decisions.

1.1. Profit Function

Following the classic inventory management results, the expected profit is

$$E(Profit) = p.min(Q, D) - c.Q - h.(Q - D)^+ - b.(D - Q)^+.$$

- $min(Q, D)$ represents the number of units sold which is the minimum of the order quantity Q and the actual demand D.
- $(Q - D)^+$ represents the excess inventory, which is the number of units that were not sold. The + notation denotes the positive part, i.e., it equals $Q - D$ when $Q \geq D$, otherwise zero.
- $(D - Q)^+$ represents the shortfall, which is the number of units that were demanded but not available.

Using standard stochastic optimization one can define the optimal critical ratio as:

$$CR^* = \frac{p - c}{p - c + b}.$$

- Where $p - c$ represents understock (underestimated) cost.

The above ratio represents the proportion of demand that should be met by the order quantity, balancing the cost of overstocking and understocking. Next, the optimal order quantity, Q^*, is found by solving for the demand distribution's quantile corresponding to the critical ratio:

$$P(D \leq Q^*) = CR^* \rightarrow Q^* = F^{-1}(CR^*)$$

Therefore, finding the demand distribution is critical in finding the optimal order quantity. Traditionally, it is assumed that demand follows a known distribution like Normal distribution. However, this is not a fair assumption in practice because demand has complex shape and trend. To find the underlying demand distribution give a dataset, we propose using VAEs. Variational Autoencoders (VAEs) can be effectively utilized to model and analyze demand distributions by leveraging their capacity to learn latent representations from data. By training on historical demand data, the encoder network of a VAE extracts latent variables that capture underlying patterns and variability in demand. The decoder network then reconstructs the demand data from these latent variables, effectively approximating the data's probabilistic distribution. Once trained, the generative capabilities of the VAE allow sampling from the learned latent space to produce synthetic demand scenarios. This enables practitioners to explore a range of plausible demand outcomes under various conditions, account for uncertainty, and evaluate strategies for robust decision-making. By

incorporating features such as time, location, and product characteristics into the input, the VAE can generate diverse demand scenarios tailored to specific contexts, making it a powerful tool for demand forecasting and planning in dynamic environments.

We borrow our VAE model from Golchin & Rekabdar (2024). The VAE assumes that data **x** is generated from a latent variable **z** through a generative process. It assumes the prior distribution over latent variables is: $p(z) = N(z; 0,1)$, which is a multivariate standard normal distribution. Therefore, the likelihood of the data given the latent variable is:

$$p(z) = Decoder\ (z)$$

where the decoder network parameterizes a likelihood distribution, often a Gaussian

$$p(z) = N\left(x;\ \mu_\theta(z), \sigma_\theta^2(z)\right)$$

Next, note that the posterior $p(z)$ is generally interactable, so the VAE approximates it using a variational distribution $q_\phi(z|x)$, parametrized by an encoder network

$$q_\phi(x) = N\left(z;\ \mu_\phi(x), \sigma_\phi^2(x)\right)$$

The objective in VAE is to maximize the Evidence Lower Bound (ELBO), which is a variational approximation to the log marginal likelihood $log\ log\ p(x)$. ELBO is defined as

$$L(x;\ \theta, \phi) = E_{q_\phi(z|x)}[log\ log\ p_\theta(x|z)\ ] - KL\big(q_\phi(x)||p(z)\big).$$

The first term in the above expression is the Reconstruction Loss, which encourages the decoder to accurately reconstruct **x** given **z**, and the second term is KL Divergence, which regularizes $q_\phi(x)$ to stay close to the prior $p(z)$.

1.2. Research Methodology

This study employs a hybrid methodology that integrates the classical Newsvendor model with a Variational Autoencoder (VAE) to address the challenge of determining optimal inventory quantities under demand uncertainty. Traditional approaches often rely on assumptions of specific probability distributions, which may not accurately reflect the complexities and non-linearities of real-world demand. To overcome this limitation, we leverage the power of VAEs to learn complex demand patterns from historical data and generate more robust and adaptive order quantity recommendations. This approach allows us to capture a wider range of uncertainties, including supply chain disruptions, logistical bottlenecks, and demand shocks, ultimately leading to more informed inventory decisions and improved profitability.

1.3. Research Design

The study adopts a simulation-based comparative framework to evaluate the effectiveness of a VAE-generated demand distribution against a traditional Normal distribution in the Newsvendor context. The workflow comprises:

Demand Data Generation: Synthesize demand datasets using Normal distributions with controlled noise to simulate real-world variability.
VAE Training: Train a VAE to learn latent representations of demand patterns, enabling the generation of synthetic demand samples.
Optimal Order Quantity Calculation: Compute the critical ratio from Newsvendor cost parameters and derive the optimal order quantity (Q^*) using percentile-based methods for both classical and VAE-generated distributions.
Profit Comparison: Evaluate expected profits under both demand models to quantify the VAE's added value.

2. Numerical Experiment

A baseline demand dataset is generated from a Normal distribution, $N(\mu_D, \sigma_D)$, augmented with Gaussian noise, $N(\mu_e, \sigma_e)$, where:

μ_D	σ_D	μ_e	σ_e

$\in \{1000,5000,10000\}$	$\in \{50,200,1000\}$	$\in \{100,500,1000\}$	$\in \{25,100,500\}$

In summary, we study four scenarios:

1. **Varying Input Size:** Testing with different input sizes helps assess if the VAE can handle varying amounts of data and still capture the underlying demand distribution.
2. **Changing the Mean:** Shifting the mean of the input data helps evaluate if the VAE can adapt to changes in the average demand level.
3. **Adjusting Noise Levels:** Adding varying levels of noise to the input data tests the VAE's robustness to noisy or uncertain demand signals.
4. **Introducing Outliers:** Adding outliers to the input data tests the VAE's ability to handle extreme or unusual demand values.

The objective is to mimic unobserved disruptions (e.g., logistics delays). The combined distribution ensures non-negativity and reflects realistic demand fluctuations. Data normalization is avoided to preserve interpretability of the VAE's output. Next, the cost parameters of the problem are as follows. The cost parameters are set so that they mimic most of the daily items found in grocery stores (e.g., chips, bread, etc.)

- Purchase cost (c): 5
- Selling price (p): 12
- Backorder penalty (b): 3
- Holding cost (h): 2

Following the cost parameters, the underage cost (C_u) which represents the lost profit and penalty for insufficient inventory is $C_u = p - c + b = 12 - 5 + 3 = 10$, and the overage cost ($C_o$) which reflects the net loss from excess inventory is $C_o = c - h = 5 - 2 = 3$. Therefore, the critical ratio, which determines the optimal service level to minimize expected cost is:

$$CR^* = \frac{C_u}{C_u + C_o} = \frac{10}{10 + 3} \approx 0.769$$

Assuming demand *D* follows a normal distribution $N(\mu, \sigma^2)$, the optimal order quantity Q^* is calculated using the inverse cumulative distribution function (CDF):

$$Q^* = \varphi^{-1}(\alpha, \mu, \sigma),$$

Where $\varphi^{-1}$is the inverse CDF (e.g., norm.ppf in Python). For a sample size of 1,000, these yields Q^* that aligns with the critical ratio. Next, the expected profit is computed using the piecewise function:

$$profit = \{pD - cQ - h(Q - D), if\ D \leq Q, pQ - cQ - b(D - Q), if\ D > Q.$$

This is vectorized as np, where$(D \leq Q, pD - cQ - h(Q - D), pQ - cQ - b(D - Q))$, where *D* is a vector of demand samples.

Next, one can observe that while the normal distribution provides a tractable solution, it fails to capture asymmetric, multimodal, or context-dependent demand patterns caused by real-world disruptions (e.g., logistics delays, supplier bottlenecks). To address this, we propose augmenting the Newsvendor model with **Variational Autoencoders (VAEs)**, which learn latent representations of demand variability from data infused with synthetic noise. VAEs are generative models that encode input data into a latent space and decode it back, enabling the generation of new samples that preserve the statistical properties of the training data. Here, the VAE is trained to model demand distributions that account for unobserved uncertainties.

To train the VAE, 1,000 samples are generated with varying μ and σ to simulate diverse demand scenarios as mentioned in Table 1. Our VAE architecture is as follows:

- **Encoder**: A neural network that maps input demand samples $\boldsymbol{x}$ to latent variables $\boldsymbol{z}$
  - Layers: Two dense layers (64 neurons, ReLU activation).
  - Output: Parameters of the latent distribution ($\mu_z, log\ (\sigma_z)$.

- **Decoder:** Reconstructs demand samples from latent variables z
  - Layers: Two dense layers (64 neurons, ReLU activation).
  - Output: Reconstructed demand $\underline{x}$
- **Implementation detail:**
- VAE Training: Adam optimizer (learning rate 10− 3), 500 epochs, batch size 64.
- Code Libraries: TensorFlow/Keras for VAEs, NumPy for numerical operations, SciPy for statistical functions.

The loss function combines reconstruction loss (mean squared error) and KL divergence to regularize the latent space. Finally, after training, the VAE generates 1,000 demand samples by sampling z from the latent prior $\boldsymbol{p(z)}$ and decoding it. These samples form a non-parametric demand distribution that reflect learned uncertainties. Next, as the generated output data distribution is not a regular probability distribution, we would not be able to estimate the cumulative probability distribution from VAE Output. Therefore, we used the proportion-based estimate to calculate the optimal demand quantity using the following procedure:

- Sort the generated demand samples **D**, from the ***VAE***
- Compute the index corresponding to the critical ratio that was calculated for our scenario using the underage cost and the overage cost.
- The profit for each demand sample is calculated using the same piecewise function as in the classical model. The expected profit from the VAE-generated demand is compared to that from the normal distribution.

Numerical Results

The histogram in Figure 2 represents the demand distribution generated using a Variational Autoencoder (VAE) trained on simulated data. Initially, demand values were sampled from a normal distribution with added noise to introduce variability according to Table 1. The VAE was then employed to learn the underlying structure of the demand data, capturing complex, potentially non-Gaussian patterns. The resulting distribution, as seen in Figure 1, exhibits a right-skewed shape, indicating that higher demand values occur less frequently, while lower demand values are more common. This highlights the advantage of using VAEs over traditional parametric approaches, as they can effectively learn and reconstruct demand distributions without assuming a predefined shape. By leveraging this data-driven approach, inventory management decisions in the Newsvendor model can be improved, allowing for more accurate order quantity determinations and reduced risks of stockouts or overstocking.

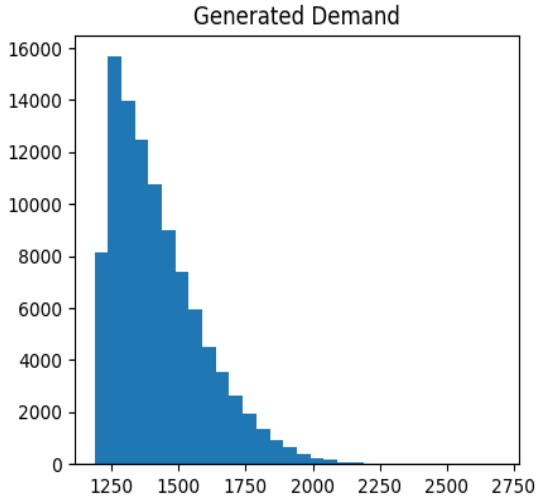

Figure 2. Demand Distribution generated using VAEs

The comparison between the two profit analyses—one derived from the Variational Autoencoder (VAE) generated demand distribution (i.e., Figure 3) and the other based on a traditional normal distribution (i.e., Figure 4)—

demonstrates the superiority of the VAE approach. The VAE-generated profit exhibits a higher average value and greater variability, suggesting that it captures more nuanced demand patterns that a normal distribution fails to represent. Notably, the maximum profit achieved using the VAE is significantly higher than that of the normal distribution, implying that the learned distribution better reflects real-world demand fluctuations. Additionally, the red trend line in the VAE-based graph maintains a consistently higher position than its normal distribution counterpart, reinforcing the idea that using VAE for demand estimation leads to more accurate and profitable inventory decisions in the newsvendor setting.

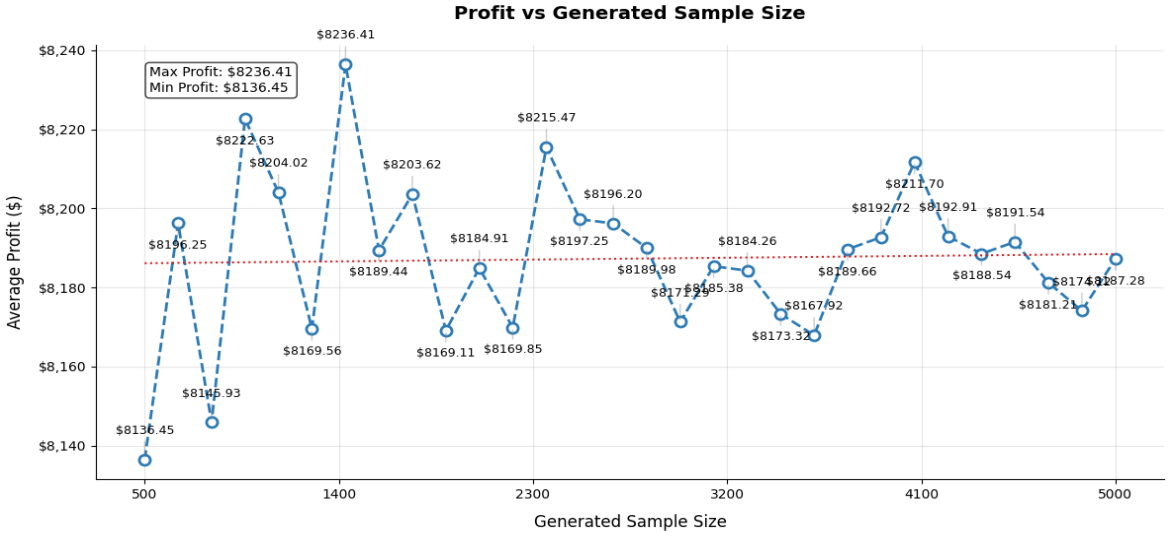

Figure 3. Profits Based on the VAE generated demand distribution

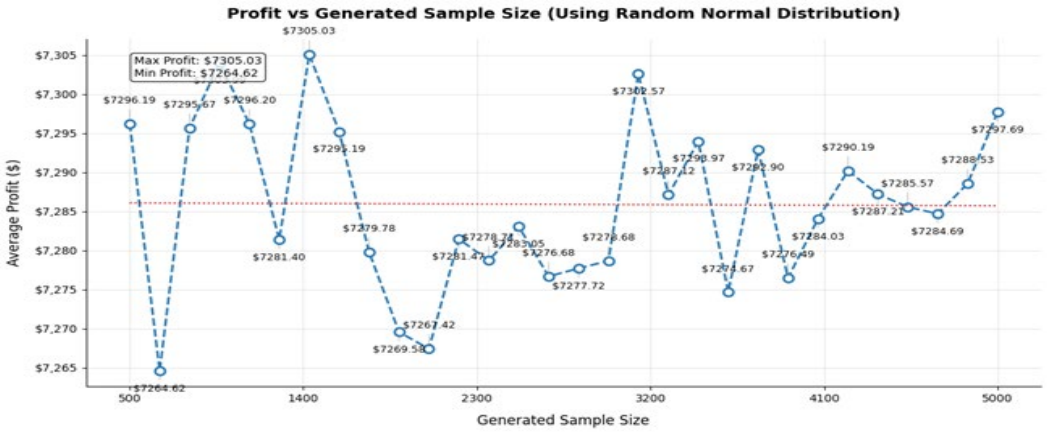

Figure 4. Profits Assuming Normal distribution for the demand.

Figures 5 and 6 compare the profit versus the sample mean. In Figure 5, we show the profit versus the sample mean assuming normal distribution. We find that all the profits lie on the red-dashed line because the underlying distribution is Normal. In Figure 6, we illustrate the profit versus the sample mean generated from the VAE. We observe here that the profits calculated assuming the VAE are generally higher than the profits calculated assuming Normal distribution.

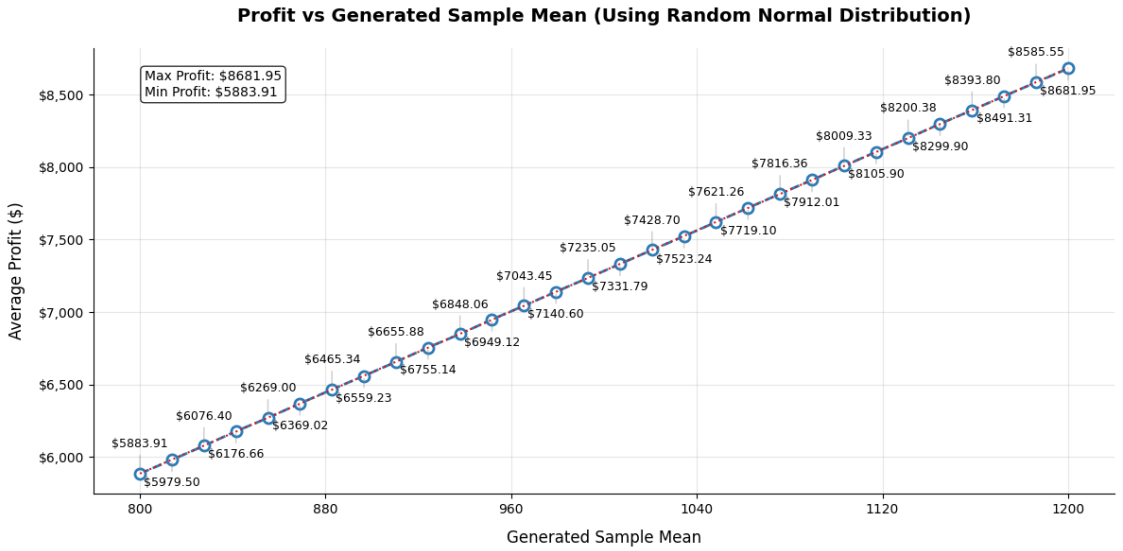

Figure 5. Profits vs. Sample Mean assuming Normal Distribution

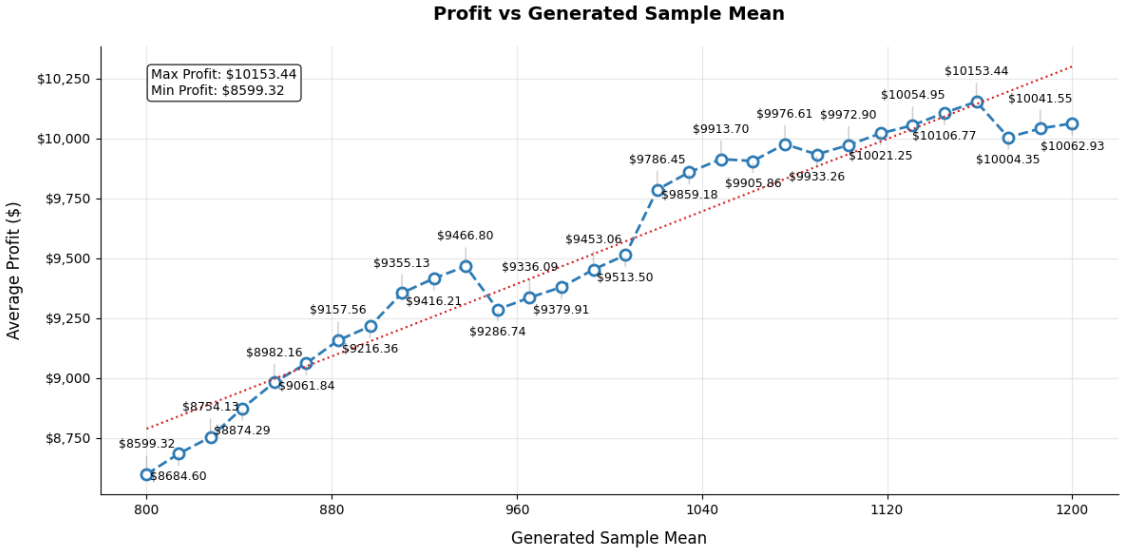

Figure 6. Profit vs. Sample Mean Generated from the VAE

4.Conclusion

The comparison between the two approaches—one using a Variational Autoencoder (VAE) to estimate the demand distribution and the other assuming a normal distribution—demonstrates the clear advantages of the VAE-based method. The profit trends in the VAE-generated model exhibit greater stability and higher average profitability compared to the normal distribution approach, which shows more fluctuations and lower overall profits. This suggests that the normal distribution assumption may oversimplify the underlying demand structure, failing to capture complex variations and real-world uncertainties. In contrast, the VAE effectively learns intricate demand patterns from data, allowing for a more precise estimation of the true demand distribution. This, in turn, leads to more informed inventory decisions and improved profitability in the newsvendor setting.

Future research can explore further enhancements to the VAE framework, such as incorporating additional external factors like seasonality, economic trends, and competitor behavior into the model to refine demand estimation.

Additionally, integrating reinforcement learning with VAEs could provide adaptive decision-making strategies that dynamically adjust inventory levels based on real-time data. Investigating the robustness of VAEs in handling sparse or limited data scenarios is another promising direction, as real-world demand data may often be incomplete or noisy. Overall, our findings highlight the potential of deep learning models in supply chain optimization, paving the way for more intelligent and data-driven inventory management solutions.

References

Araújo, A. P. F., Costa, A. P. C. S., and Scavarda, L. F., Data-driven solutions for the Newsvendor problem: A systematic literature review, Advances in Production Management Systems, vol. 23, no. 2, pp. 85-97, 2022.

Feng, Y., Cai, X., and Chen, X., A data-driven Newsvendor problem: Insights from machine learning, European Journal of Operational Research, vol. 305, no. 1, pp. 248-263, 2023

Golchin, B., and Rekabdar, B., Anomaly detection in time series data using reinforcement learning, variational autoencoder, and active learning, Proceedings of the 2024 Conference on AI, Science, Engineering, and Technology (AIxSET), pp. 1-8, Laguna Hills, CA, September 30 – October 2, 2024.

Golchin, B., and Riahi, N., Emotion detection in Twitter messages using combination of long short-term memory and convolutional deep neural networks, International Journal of Computer and Information Engineering, vol. 15, no. 9, pp. 578-585, 2021.

Hao, Ruoyu., Streamlining SCM: Integrating Demand Forecasting and Inventory Optimization, In 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024), pp. 550-558. Atlantis Press, 2024.

Lim, B., and Zohren, S., Time-series forecasting with deep learning: a survey, Philosophical Transactions of the Royal Society A, vol. 379, no. 2194, pp. 20200209, 2021.

Mohamadi, N., Niaki, S. T. A., Taher, M., and Shavandi, A., An application of deep reinforcement learning and vendor-managed inventory in perishable supply chain management, Engineering Applications of Artificial Intelligence, vol. 127, pp. 107403, 2024.

Müller, S., Reijers, H. A., and Borghoff, U. M., A data-driven Newsvendor problem: From data to decision, International Journal of Production Economics, vol. 193, pp. 351-362, 2017.

Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E. A., and Lima Netto, S., Variational autoencoder. In Variational Methods for Machine Learning with Applications to Deep Networks, (pp. 111-149). Cham: Springer International Publishing, 2021.

Praveen, K. B., Kumar, P., Prateek, J., Pragathi, G., and Madhuri, J., Inventory management using machine learning, International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 6, pp. 866-869, 2020.

Qin, Y., Wang, R., Vakharia, A. J., Chen, Y., & Seref, M. M., The newsvendor problem: Review and directions for future research, European Journal of Operational Research, vol. 213, no. 2, pp. 361-374, 2011.

Rossi, R., Prestwich, S., Tarim, S. A., and Hnich, B., Confidence-based optimisation for the newsvendor problem under binomial, Poisson and exponential demand, European Journal of Operational Research, vol. 239, no. 3, pp. 674-684, 2014.

Saghafian, S. and Brian, T., The newsvendor under demand ambiguity: Combining data with moment and tail information, Operations Research, vol. 64, no.1, pp. 167-185, 2016.

Simchi-Levi, D., and Zhao, Y., The big data Newsvendor: Practical insights from machine learning analysis, Management Science, vol. 67, no. 3, pp. 1436-1452, 2021.

Vázquez-García, C., Martínez-Murcia, F. J., Segovia Román, F., and Górriz Sáez, J. M., Tutorial: VAE as an inference paradigm for neuroimaging, arXiv preprint, arXiv:2501.08009, 2025.

Khlie, K., Benmamoun, Z., Fethallah, W., and Jebbor, I., Leveraging variational autoencoders and recurrent neural networks for demand forecasting in supply chain management: A case study, Journal of Infrastructure, Policy and Development, vol. 8, no. 8, pp. 6639, 2024.

Biographies

Mohammad E. Arbabian is an Assistant Professor at the Pamplin School of Business, University of Portland, Portland, Oregon. He earned his B.S. in Industrial Engineering focusing on Technology Management at Iran University of Science and Technology, Tehran, Iran. He earned his M.S. in Industrial Engineering focusing on System Analysis at K.N Toosi, Tehran, Iran. He then earned his PhD in Operations Management focusing on Emerging Technologies at University of Washington, Seattle, Washington. He has published papers in top journals in Operations Management including *Manufacturing and Service Operations Management*, *European Journal of Operations Research*, *International Journal of Production of Economics*, and *Operations Management Research*. He also has published conference papers. His papers have garnered extensive citations. His primary focus is on Additive Manufacturing, Inventory Management, Emerging Technologies, Optimization and Supply Chain Management. He is a member of POM, INFORMS, MSOM and DSI. Relating to his Industry Experience, he is the cofounder of RayPars consulting company.

Dr. Naga Vemprala is an Assistant Professor of Operations and Technology Management at the University of Portland. His research focuses on the intersection of social media, data analytics, machine learning, and natural language processing. Dr. Vemprala has published in numerous prestigious academic journals, including the International Journal of Information Management, Decision Sciences Journal, ACM Transactions in Management

Information Systems, Information Systems Frontiers, and American Behavioral Scientist. He has also presented his work at leading international conferences and workshops such as the Bright Internet Global Summit, AMCIS, ECIS, and HICSS. Dr. Vemprala earned his Ph.D. in Information Technology from the University of Texas at San Antonio and holds a bachelor's degree in Electrical and Electronics Engineering from Andhra University. He has over a decade of experience in the IT service industry, working as a software developer and data analyst for multinational companies.

Dr. Naveen Gudigantala is the Silicon Valley Distinguished Professor of Operations and Technology Management at the University of Portland. His research examines consumer and firm behavior in online environments, the application of artificial intelligence and machine learning in organizations, and the creation of value through business analytics and AI. His research interests also include supply chain ethics, intellectual property rights, optimizing machine learning pipelines, and information systems education. Dr. Gudigantala has published his work in leading information systems journals, including Communications of the AIS, Decision Support Systems, International Journal of Information Systems, and the Journal of Information Systems Education. He is an active member of the IS academic community, serving as a reviewer for numerous journals and conferences. Dr. Gudigantala earned his Ph.D. from Texas Tech University.