Received: 16 January 2026 | Revised: 22 April 2026 | Accepted: 20 May 2026 | Published online: 20 July 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–16
DOI: 10.47852/bonviewJDSIS62029121

BON VIEW PUBLISHING

AI-Assisted Synthetic Dataset Generation and Validation for Demand Forecasting in the Context of Mass Customization

Nouhaila El Assad[1,*], Kawtar El Haouti[1], Soumaya Zayrit[1], Salah-eddine Mokhlis[1], Mohamed Rhouzali[1] and Najat Messaoudi[1]

[1] *Laboratory of Electrical and Industrial Engineering, Information Processing, Computing, Logistics—Renewable Energies and Systems Dynamics, Hassan II University Casablanca, Morocco*

Abstract: Mass customization emerged as a strategic response to the increased consumer interest in personalized products, combining the efficiency of mass production with the flexibility of individual customization. Yet, this manufacturing paradigm introduced significant complexities in demand forecasting, where traditional methods struggled to account for the high variability in customer preferences. The lack of accessible, high-quality datasets that capture real customization behavior makes this challenge more difficult. To address this gap, this study develops and validates a large-scale synthetic dataset simulating 500,000 laptop sales. Each sale corresponded to a customized product configuration tailored to an individual customer profile, thereby reflecting the dynamics of real-world mass customization contexts. The dataset was generated using a structured rule-based framework, supported by artificial intelligence (AI)-assisted reasoning and grounded in domain knowledge and manufacturing literature. This framework was designed to reproduce realistic customer behavior, configuration patterns, and seasonal demand fluctuations. As a result, the generated data captured heterogeneous customer profiles and varied product configurations. Its realism and coherence were assessed through statistical distribution analysis, logical consistency checks, and seasonal pattern verification. The validation results provide a reliable basis for evaluating and benchmarking demand forecasting models. Overall, this work shows that combining rule-based design with AI-assisted reasoning is effective for the generation of realistic synthetic datasets for demand forecasting in mass customization contexts where real data is unavailable.

Keywords: mass customization, dataset generation, dataset validation, demand forecasting, Industry 4.0

1. Introduction

Over the past century, industries have undergone several major transformations. In the early stages, craft production was dominant, with skilled artisans producing goods for individual customers. This model was eventually replaced by mass production, which focused on efficiency, standardization, and economies of scale. This transition made products more affordable and accessible, but at the expense of variety and personalization.

Mass customization is a manufacturing approach that originated in the late 1980s due to increasing customer demand for variety. This approach merged the cost efficiencies of mass production with the ability to personalize products. By leveraging modular design and advanced production technologies, companies could offer customers the ability to customize aspects of their purchases, such as laptop specifications or car colors, while maintaining economies of scale [1–3].

Table 1 highlights some of the core differences between mass production and mass customization.

The emergence of Industry 4.0 has accelerated to a great extent the shift toward mass customization. Technologies such as artificial intelligence (AI), Internet of Things (IoT), and cyber-physical systems enable manufacturers to collect real-time data, automate production processes, and respond with more flexibility to customer demands. Studies have also highlighted the importance of Industry 4.0 readiness and maturity models in supporting this transition, in particular within manufacturing contexts [4, 5]. This has allowed businesses across electronics, automotive, and fashion sectors to deliver customized products without sacrificing efficiency [2].

Figure 1 [6] illustrates the evolution of production modes from craft production to mass personalization, while taking into consideration the product variety and productivity of each mode.

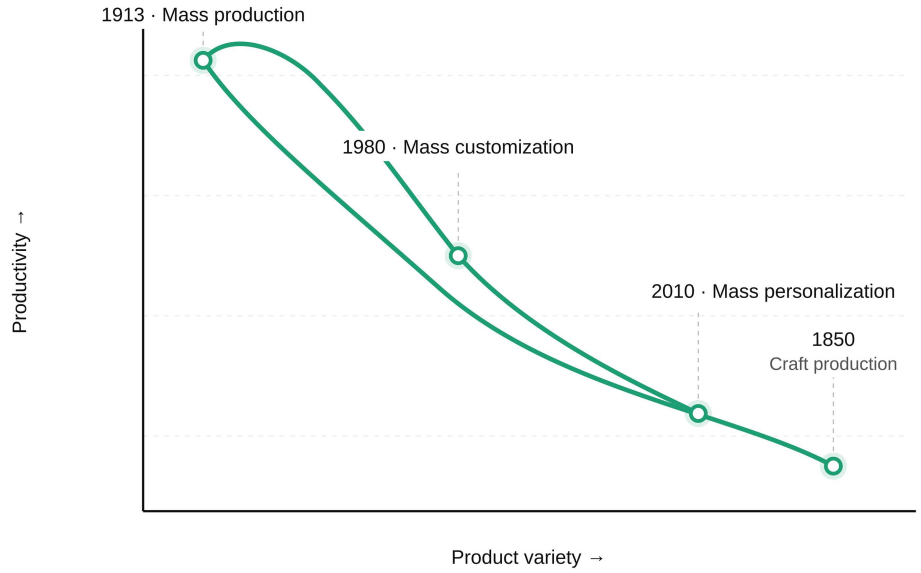
In the context of this study, product variety refers to the number of possible product configurations that can be created within a product family based on different combinations of product components or attributes. In mass customization environments, product variety does not necessarily imply different product categories. Instead, product variety is the large number of variants

***Corresponding author:** Nouhaila El Assad, Laboratory of Electrical and Industrial Engineering, Information Processing, Computing, Logistics—Renewable Energies and Systems Dynamics, Hassan II University Casablanca, Morocco. Email: nouhaila.elassad-etu@etu.univh2c.ma

Table 1
Mass production vs mass customization

Aspect	Mass production	Mass customization
Demand uncertainty	Stable demand forecasting based on historical data	Higher uncertainty due to individual customer preferences
Production strategy	Push strategy: goods produced based on forecasts and stocked	Pull strategy: production starts after an order is received
Flexibility and responsiveness	Low flexibility; optimized for efficiency	High flexibility to accommodate customer-specific variations
Technology dependence	Traditional assembly lines with minimal real-time adaptation	Uses advanced technologies (AI, additive manufacturing, modular design)
Cost and efficiency	Economies of scale, lower unit costs, risk of overproduction	Balances efficiency with variety, higher per-unit cost due to customization
Lead time and inventory	Maintains inventory for quick delivery but risks unsold stock	Reduces inventory but may lead to longer lead times

Figure 1
The evolution of production modes



of a specific product [7]. In the laptop industry, for example, product variety is a result of combinations of different configurable product components or attributes such as processors, graphics cards, memory capacity, storage type, and display size. The more configurable attributes or components, the greater the number of possible product configurations, thus increasing operational complexity and creating additional challenges for demand forecasting and manufacturing planning [8].

As product variety expands, demand data become increasingly fragmented across a large number of possible product configurations, making it difficult to obtain structured datasets that accurately represent demand behavior in mass customization environments.

This growing flexibility benefits customers but poses enormous operational challenges for companies in the light of supply chain management. Traditional push-based production mode, where companies manufacture in advance based on a stable demand forecast, is incompatible with environments characterized by real-time and individualized orders. Mass customization systems rely on pull-based production, where products are assembled only after a customer order is received while components remain available for configuration. This is an area that also calls for a high level of agility and visibility [9].

In today's volatile markets, where product life cycles have shortened and customer expectations shift in a rapid way, accurate demand forecasting has become critical. It affects procurement, inventory control, and production scheduling, determining whether companies can respond fast without excessive costs or inventory buildup [2, 7, 10].

In mass production, demand usually follows recognizable patterns and historical trends. Mass customization creates a different problem because it introduces high levels of demand uncertainty. Each order may be unique and can combine product features in a specific way, which increases the number of possible configurations. This makes future demand harder to estimate with traditional forecasting methods. Changing customer preferences and complex supplier dependencies add another layer of uncertainty to the forecasting task [7].

Research in this area also faces a practical data problem. The scarcity of open-access datasets that describe demand behavior in mass customization with sufficient precision is still considered a major problem. Most studies on demand forecasting in this context use proprietary or case-specific industrial data. These data are not publicly available, which limits reproducibility and makes it difficult to compare forecasting models across studies [11, 12]. Synthetic data generation has already been used to

address data scarcity and confidentiality issues in manufacturing systems [13]; however, no open-access datasets have been specifically designed to represent demand behavior in mass customization environments. While some industrial datasets exist, they remain confidential and not accessible for direct comparison or reproducibility. This makes the use of a validated generated dataset necessary in this study.

The absence of accessible data limits the systematic development and evaluation of forecasting methods for highly configurable products. Addressing this gap is therefore important for supporting data-driven forecasting analysis, procurement planning, and supply chain decision-making in highly configurable product environments.

This study aims to develop and validate a synthetic dataset that represents demand behavior in a mass customization context. The dataset is designed to realistically reproduce customer profiles, configurable laptop attributes, and seasonal purchasing patterns observed in the laptop industry. By providing a structured and validated dataset, this work supports the development and benchmarking of demand forecasting models for mass customization environments.

To achieve this goal, the research analyzes existing dataset limitations, designs an AI-assisted data generation framework, integrates realistic seasonal and behavioral dynamics, and validates the dataset through statistical and logical consistency analyses to ensure representativeness and reliability.

2. Literature Review

2.1. Demand in mass production vs in mass customization

Understanding demand behavior across manufacturing paradigms helps clarify why forecasting is difficult in mass customization. In mass production, companies usually deal with stable and predictable demand for standardized products. They rely on historical sales data and time-series models to forecast sales volumes and to produce goods in large quantities ahead of anticipated demand.

In most cases, forecasts are aggregated, adjusted over seasons, and projected over extended planning horizons [1].

Mass customization introduces greater variability and complexity. Unlike mass production, where demand is concentrated on a limited number of standardized products, mass customization distributes demand across a large number of product configurations. Each configuration may appear only rarely in historical data, resulting in fragmented demand patterns that are significantly more difficult to analyze and predict [7].

Mass customization refers to the ability of manufacturers to offer products tailored to individual customer preferences while maintaining efficiency levels close to those of mass production. This is typically achieved through configurable product architectures and modular design strategies that allow different product variants to be assembled from a set of predefined components [8, 14].

Customers engage in product design, selecting features such as hardware configurations or esthetic preferences. To manage this increasing variety, many manufacturers rely on modular design and product platform strategies that allow companies to generate multiple product variants while controlling internal complexity and production costs [15]. This fragments demand across a much wider range of product combinations, many of which are ordered only a few times, or even only once. As a result, conventional grouping and forecasting methods become inadequate [7].

As product variety increases, managing the complexity associated with numerous product configurations becomes a significant challenge for manufacturing systems [16].

2.2. Mass customization and forecasting challenges

The underlying production logic difference is also important in terms of the implications that it has on forecasting accuracy and supply chain planning. Traditional forecasting methods designed for mass production, in which demand tends to be stable and aggregated, struggle to capture the dynamic, personalized nature of mass customization [2].

Similar challenges have been observed in make-to-order manufacturing systems, where demand often appears as intermittent and highly variable across product configurations, making conventional forecasting approaches less effective [17].

The challenge stems from extreme product diversity. A single product line might branch into hundreds or even thousands of possible configurations. In practice, this variety is often managed through modular product platforms that enable companies to generate multiple product variants from a common set of components [8]. This diversity makes it very difficult to apply the same models used in mass production, where demand is smoother and more consistent [9].

Market dynamism and technology innovation cycles make it even more difficult to have accurate forecasting. Upgrading many products is necessary, and customer behavior becomes less predictable. At the same time, product life cycles have shortened, but procurement and assembly cycle times remain long, creating a gap between demand and supply prospects. Supply bases have become more volatile, leading to longer cycle times and limited capacity, and small margins of error in forecasting have ensured operational disruptions in both inventory surpluses and shortages [15].

2.3. Synthetic data in manufacturing and forecasting

Data granularity presents an additional challenge. Effective mass customization forecasting requires understanding not only what was sold but also who purchased it, their motivations, and how preferences mapped to specific product features. Few companies maintain such detailed data in usable formats, and researchers don't gain access to it [2, 18]. Traditional time-series and regression models struggle with the high-dimensional, sparse datasets characteristic of mass customization. This has driven calls for more adaptive, real-time, and data-driven forecasting approaches tailored to this context [19, 20].

The growing emphasis on data-driven models has heightened the need for high-quality datasets. In manufacturing and supply chain management, synthetic data generation offers a solution when real operational data are confidential, limited, or unavailable, allowing simulation of realistic customer behaviors, product configurations, and supply chain dynamics without compromising sensitive information.

Recent studies have explored synthetic data in manufacturing contexts, proposing frameworks to simulate operations to train machine learning models [11, 12]. Simulation-based approaches have also been used to generate synthetic manufacturing datasets for machine learning applications [21]. In parallel, research on synthetic time-series generation has shown the importance of evaluating the statistical fidelity of synthetic datasets used in machine learning [22]. Similar approaches have proven valuable in healthcare, retail, and energy systems [23]. Yet few synthetic datasets target mass customization, and even fewer are open-access or

validated. Published datasets often remain proprietary or lack detailed documentation.

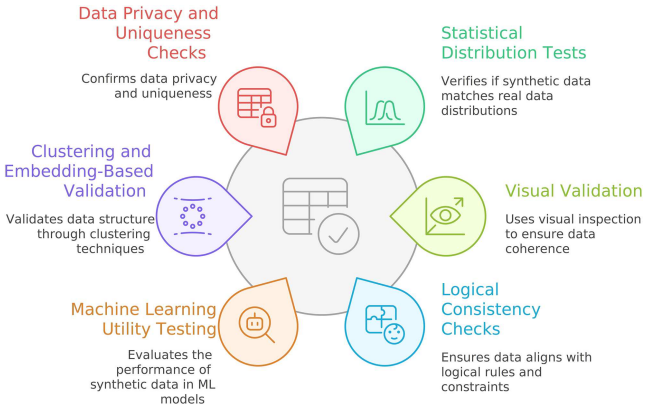
This creates a significant obstacle for demand forecasting research in mass customization. Most studies depend on either confidential company data or small-scale academic simulations that fail to capture the full complexity of customer preferences and product configurations. Without standardized, public datasets, comparing models, validating approaches, and reproducing experiments become difficult [18, 24]. Validation methods are underdeveloped, and assessments tend to be visual or informal rather than employing structured techniques to evaluate distribution fidelity, customer-product logic, or temporal patterns.

These limitations stem from several factors such as industrial datasets that remain protected by confidentiality agreements, generating realistic synthetic data requiring substantial domain expertise and computational resources, and no consensus exists on appropriate validation metrics for assessing realism and statistical coherence. It's crucial and very important for advancing reproducible forecasting research to address these challenges with validated open-access datasets. It also helps strengthen the link between academic work and industrial practice.

Synthetic dataset validation checks whether the generated data is representative, accurate, and suitable for further use. This step is especially important when the dataset supports forecasting, simulation, or decision support models. In mass customization, demand results from the interaction between customer preferences, modular product configurations, and temporal variation. Validation therefore helps confirm that the generated data is statistically realistic and consistent with domain logic.

The literature proposes several methods for assessing synthetic data validity, as shown in Figure 2. These methods include statistical distribution tests, visual validation, logical consistency checks, machine learning utility testing, clustering and embedding-based validation, and privacy and uniqueness checks. Each method examines synthetic data quality from a different angle, depending on the intended use of the dataset.

Figure 2
Validation methods for synthetic datasets



Statistical tests measure how closely the generated data follows the expected distributions. Machine learning evaluations assess whether the data can support useful predictive models. Logical consistency checks verify that the generated records respect domain rules. Clustering and embedding-based methods examine structural similarity between datasets. Visual validation helps researchers inspect patterns qualitatively, while privacy-related checks assess uniqueness and disclosure risk. Table 2 summarizes the validation approaches, including their strengths, limitations, and applications [13, 23, 25–27].

Despite the existence of several validation strategies, applying these approaches to mass customization datasets remains challenging due to the high dimensionality of product configurations and the complex relationships between customer profiles and product attributes.

3. Materials and Methods

3.1. Type of study

This is a computational and simulation-based study that aims to create a synthetic dataset for demand forecasting in a mass customization context. This study is non-observational, since it does not rely on data collected from real customers or companies. All the records are synthetically generated using a rule- based framework assisted by AI reasoning. The objective is to reproduce realistic purchasing behavior, variability in configurations, and seasonal demand dynamics in the laptop industry in mass customization contexts. This is a quantitative method involving probabilistic modeling, logical reasoning, and validation tests to promote consistency and analytic validity in the generated dataset.

3.2. Universe of the study

The universe represents the electronic industry, specifically the laptop industry with modular products and customization options. It includes the full variety of all possible different orders of laptops that might occur in a given year in a mass customization environment.

3.3. Sample

The sample includes around 500,000 synthetic laptop sales transactions generated to reproduce this universe.

Each record represents:

1) A customer profile (segment, age, income, gender);

Table 2
Overview of synthetic data validation approaches: methods, strengths, weaknesses, and use cases

Validity approach	Example methods	Strengths	Weaknesses	Common usage contexts
Statistical (fidelity)	Univariate tests (K-S, chi-square), distribution distances (Wasserstein, Jensen-Shannon Divergence (JSD)), correlation difference, two-sample tests (maximum mean discrepancy (MMD)), coverage rate	Easy to compute; interpretable; fast checks on fidelity	May miss multivariate patterns; not task-specific	General fidelity check in most studies

(*Continued*)

Table 2
(*Continued*)

Validity approach	Example methods	Strengths	Weaknesses	Common usage contexts
ML-based (predictive)	TSTR (train-on-synthetic, test-on-real), train-on-real, test-on-synthetic (TRTS), classifier Area Under the Curve (AUC) (real vs. synthetic), clustering separation	Measures predictive utility; detects distributional gaps	Results depend on model/-task choice; computational cost	ML tasks, benchmarking generators (e.g., SDGym)
Utility-based (analytic)	Propensity score MSE (pMSE), log-likelihood under real-data model, analysis replication, composite utility scores	Task-specific evaluation; reflects practical usability	Context-dependent; needs real data for baseline	Applied tasks in healthcare, finance, etc.
Privacy risk	Nearest-neighbor (Distance to Closest Record (DCR)), membership/attribute inference attacks, singling out, k-anonymity	Simulates attacks; ensures safe sharing; regulatory relevance	Metrics depend on assumptions; tension with fidelity	Healthcare, public synthetic data release
Visual/qualitative	Principal component analysis (PCA)/t-distributed stochastic neighbor embedding (t-SNE) plots, histograms, pair plots, rule violations, expert reviews	Easy to interpret; domain expert involvement; anomaly detection	Subjective; doesn't scale well; limited to few dimensions	Exploratory analysis, healthcare, interpretability

2) A customized laptop configuration (CPU, GPU, RAM, storage, screen size, battery capacity, keyboard, OS, color);
3) A timestamp within a synthetic annual sales cycle.

3.4. Sampling logic

The sample was obtained using probabilistic stratified sampling to ensure equal representation of the five defined customer segments, which are students, gamers, professionals, executives, and general consumers. The sample size was 500,000 records for obtaining strong statistical inferences with a wider coverage of various configuration combinations and close-to-realistic attribute distributions that have been observed in the laptop market.

3.5. Variables

The dataset is composed of three types of variables: customer variables, product configuration variables, and temporal variables. These variables are used to describe the characteristics of each simulated transaction:

1) Customer variables

a. Customer profile: a categorical variable representing one of five mutually exclusive consumer segments (student, gaming enthusiast, tech professional, business executive, general consumer).
b. Age category: a categorical variable indicating the buyer's age group (< 25, 25–35, 35–50, > 50). Age categories are sampled conditionally on the customer segment to reflect realistic demographic distributions.
c. Income level: a categorical variable capturing the buyer's purchasing power (low, medium, high). It is sampled conditionally on the segment and influences the likelihood of selecting premium hardware components.
d. Gender: a binary categorical variable (male/female), sampled consistently with the gender distribution typically reported in consumer electronics purchasing studies.

2) Product configuration variables

a. CPU: a categorical variable representing the central processing unit tier. The CPU is sampled conditionally on the customer segment, with high-performance tiers concentrated among gamers and professionals.
b. GPU: a categorical variable representing the graphics processing unit. GPU selection is strongly conditioned on the customer segment, with dedicated high-end GPUs concentrated in gamer and professional profiles.
c. RAM: a discrete variable representing the memory capacity of the laptop in gigabytes. The distribution is strongly segment-dependent.
d. Storage: a categorical variable indicating the storage type and capacity combination. Storage options are sampled independently of RAM, preserving configuration diversity across the dataset.
e. Screen size: a discrete variable representing the display diagonal in inches. Larger screens are more commonly selected by gamers and professionals, while students and general consumers tend to prefer compact models for portability.
f. Battery capacity: a discrete variable expressed in watt-hours (Wh), representing the battery life potential of the configured laptop. Compatibility constraints enforce that high-performance GPU and CPU configurations are paired with sufficiently large battery capacities.
g. Keyboard type: a categorical variable indicating the keyboard layout or input feature. Keyboard type is treated as an esthetic or comfort preference and is sampled with minimal segment dependency.
h. Operation system: a categorical variable indicating the pre-installed OS. OS distribution is conditioned on the customer segment.

i. Color: a categorical variable representing the chassis color option. Color is treated as an esthetic preference variable and is sampled with near-uniform probability across segments, reflecting its low functional dependency.
j. Touchscreen option: a categorical variable indicating whether the selected configuration includes a touchscreen display. Touchscreen selection is mildly conditioned on the customer segment, with executives showing slightly higher selection rates.

3) Temporal variable

a. Month (1–12), representing the synthetic annual sales cycle: a discrete integer variable representing the month of purchase within the synthetic annual sales cycle. Transactions are assigned months according to a non-uniform probability distribution designed to reproduce known seasonal demand peaks in the consumer electronics market, particularly the back-to-school period (August–September) and the holiday season (November–December).

To formalize the dataset generation process, each synthetic transaction is represented as a vector of variables describing the customer profile, product configuration, and purchase time. A transaction x_i is defined as:

$$x_i = (S_i, C_i, A_i, T_i)$$

where S_i denotes the customer segment, C_i represents the set of customer demographic attributes (customer profile, age category, income level, gender), A_i represents the product configuration attributes (CPU, GPU, RAM, storage, screen size, battery capacity, keyboard type, operating system, color, touchscreen option), and T_i denotes the temporal variable corresponding to the month of purchase.

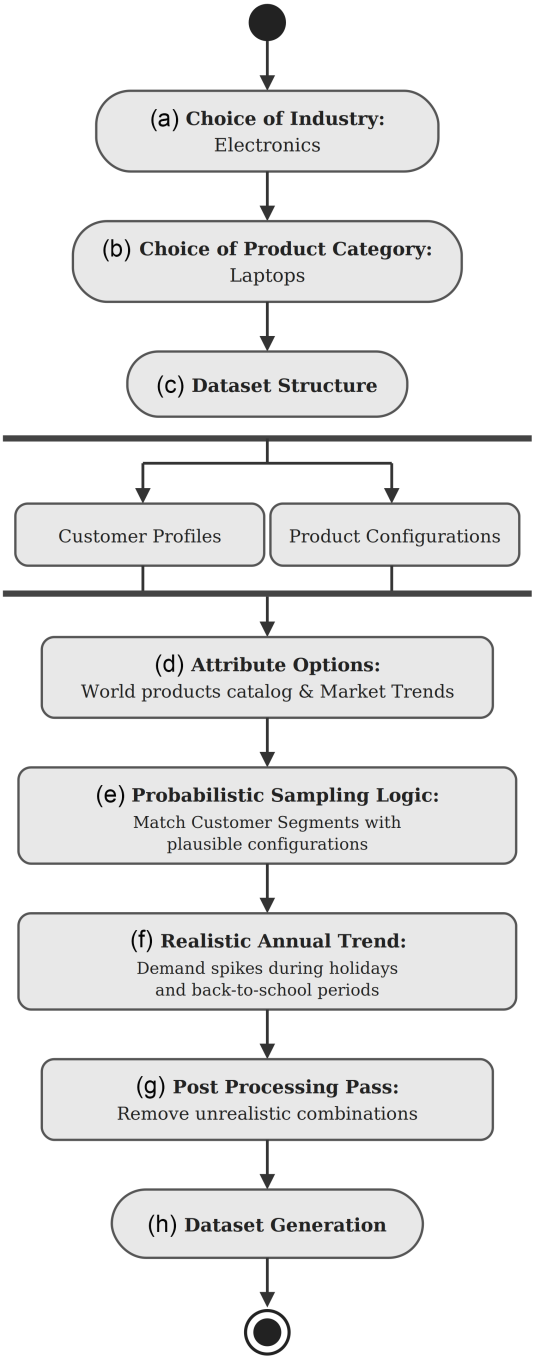
3.6. Ethical standards

This study does not involve human subjects, personal data, or any form of clinical or behavioral experimentation. The entire database used in the research is therefore synthetic and was generated, with no references to real people or confidential information. The research therefore does not pose any risks.

3.7. Synthetic dataset generation process

This subsection shows the flowchart applied to create a synthetic dataset used in this study. The process, presented in Figure 3, is a logic sequence starting from the definition of the choice of industry and concluding with the generation of the final dataset [28]. Every step helps to create realism, consistency, and diversity in customer as well as product aspects.

Figure 3
Dataset generation flowchart



3.7.1. Choice of industry: electronics

We started the process by selecting the electronics industry as the study domain. Innovation rate is high, and its product life cycle is short, which shows good customization prospects. And those characteristics make it suitable for studying forecasting challenges in configurable production contexts. Starting from a well-defined industrial perimeter, the next stages would allow a well-structured, realistic market behavior and technological diversity.

3.7.2. Choice of product category: laptops

Laptops are the target product family in the electronics industry. This category offers a variety of configurable characteristics, such as CPU, GPU, memory, display, and storage options, and well-defined customer segments, from students to professionals to gamers. Focusing on laptops created a specific and representative example where customization impacts demand variability.

3.7.3. Dataset structure

The dataset structure, built upon this foundation, was designed to combine two related dimensions:

1) Customer profiles, which include demographic and behavioral data (segment, gender, age, and income).

2) Product configurations, indicating the technical and esthetic features (processor, graphics card, RAM, storage, screen size, battery capacity, color, keyboard layout, and operating system).

This approach to dual structure set a solid theoretical basis for relating customer preferences to functional product designs, which enabled us to analyze demand in behavioral as well as configurational terms.

3.7.4. Attribute options: market catalog and trends

To populate these dimensions, the attribute options that were defined were based on global product catalogs and market trends. We assigned realistic values of each attribute to reflect the available products and technical feasibility. To conserve internal coherence, relations between attributes were maintained. The data was grounded in realist market conditions to connect abstract design logic with practical limitations.

3.7.5. Probabilistic sampling logic

After the attribute pool was established, a probabilistic sampling technique was used to link customer sectors with realistic configurations. The probability weight assigned to each segment-attribute pairing was defined to reflect preference tendencies commonly reported in the literature on consumer behavior in configurable product markets and in studies on mass customization demand heterogeneity [7, 29]. For example, gaming-oriented profiles were assigned higher probabilities for high-performance configurations, with dedicated gaming GPUs representing approximately 55–65% of their GPU choices, 16 GB or 32 GB RAM representing around 85–90% of their memory configurations, and 15-inch to 17-inch screens accounting for nearly 85–90% of their screen-size selections. Student profiles followed a different pattern. The generation framework gave them a stronger tendency toward affordable and portable laptops. Integrated or entry-level GPUs accounted for about 55–65% of their GPU choices. Low-income customers represented about 70–75% of this group, and 13-inch to 14-inch screens made up nearly 65–75% of their screen-size selections.

Business executive and tech professional profiles showed another logic. These profiles were more often associated with premium and productivity-oriented configurations. They included higher proportions of high-income customers, larger memory capacities, and professional or high-performance graphics options.

The literature did not define each probability value directly. Instead, it helped establish the expected direction of preferences across customer segments. The final probabilities were then adjusted within the generation framework to keep the dataset realistic, diverse, and internally consistent. The selected component options and their relative importance also followed publicly available trends in the global laptop market, as well as the modular logic of configurable products described in previous studies in the literature [8, 15]. Seasonal demand weights were derived from recognized commercial cycles in consumer electronics, specifically the back-to-school and end-of-year holiday periods, with August–September representing approximately one-fifth of annual demand and November–December approximately one-quarter, as documented in retail demand literature. The probability distributions were therefore not assigned arbitrarily; they were structured to reflect differentiated but nondeterministic purchasing tendencies, which are characteristic of mass customization environments. All probability weights were explicitly defined as parameters in the Python-based generation framework and are fully reproducible.

Python was used with NumPy and Scikit-learn code to build this logic, with a stratified sampling that balanced all segments across different classes, keeping diversity within each one of the classes. By allowing natural variation and occasional exceptions, we avoided patterned predictability while allowing realistic results to be obtained.

The probabilistic generation of transactions follows the joint distribution:

$$P(S, C, A, T) = P(S)\ P(C|S)\ P(A|S)\, P(T)$$

where $(P(S))$ represents the probability distribution of customer segments, $(P(C|S))$ represents the distribution of customer attributes conditioned on the segment, $(P(A|S))$ represents the distribution of product configurations conditioned on the segment, and $(P(T))$ represents the temporal probability distribution across months.

3.7.6. Realistic annual trend

To include the temporal dimension, the timestamps were given by an estimate of a synthetic annual trend. The distribution of transactions simulated seasonal characteristics commonly observed in the global consumer laptop market. In the laptop market, demand usually follows clear commercial cycles. Sales tend to rise during the back-to-school period in August and September and again during the end-of-year holiday season in November and December. These periods are commonly associated with stronger electronics purchases. Demand is generally lower in the middle of the year, when fewer major promotional events take place.

The study therefore used a temporal probability distribution that reflects these seasonal patterns without relying on sales data from a specific region. This probabilistic assignment of order dates helped produce coherent demand fluctuations across customer segments. It also allowed the generated dataset to reflect the main purchasing cycles observed in the real-world laptop industry.

3.7.7. Post-processing pass

After data generation, the dataset was checked through a post-processing step. This step removed configurations that were technically incompatible or inefficient, even if they had been produced by the stochastic sampling process, through logical consistency rules. Examples include laptops combining a high-end GPU with a very small battery or configurations with an inconsistent processor and motherboard pairing.

This correction was necessary because probability-based generation can sometimes create combinations that are statistically possible but unrealistic from a technical perspective. The post-processing step therefore helped preserve the internal coherence of the dataset and ensured that the final configurations remained compatible with real laptop design logic.

Data integrity was verified using Pandas, while Matplotlib and Seaborn were employed for visual inspections of attribute distributions and segment balance. By having this step, we ensured that the dataset remained both valid in the statistical aspect and coherent in the technical aspect.

From a formal perspective, the post-processing step applies a set of compatibility constraints to ensure the feasibility of the generated configurations. Let Ω denote the set of feasible laptop configurations satisfying all technical compatibility rules between components. A generated configuration Ai is accepted only if:

$$A_i \in \Omega$$

where Ω includes logical constraints such as compatible CPU-GPU combinations, sufficient battery capacity for high-performance components, and other hardware consistency requirements.

3.7.8. Dataset generation and export

The processed records were then consolidated into one dataset containing 500,000 synthetic transactions. Each record followed the same variable structure, with standardized categorical values and numeric ranges kept within realistic limits. Throughout the generation process, the dataset was built to avoid target leakage. The final version was exported in CSV format so it could be easily used in machine learning-based forecasting models, simulation experiments, and further analytic workflows.

Due to the absence of open-access demand forecasting datasets in the mass customization domain, ChatGPT served as an analytic assistant throughout the design process. The model was used through interactive prompts to explore plausible relationships between customer segments and laptop configuration attributes, as well as to suggest realistic attribute options based on commonly available market configurations. Besides, the model was able to help establish logical rules that would relate customer attributes to product features, identify potential edge cases, and even normalize their probability weights so that they remained balanced. It also offered reasoning support solutions for generation problems, such as over-represented specific configurations or unrealistic temporal peaks. The authors then reviewed and validated all proposed rules and distributions before implementing them in the Python-based generation framework. This step helped improve the realism and consistency of the dataset while keeping the method transparent. It also ensured that the final dataset could be reproduced from clearly defined probability rules and compatibility constraints.

Figure 4 summarizes the full generation process. It covers the main steps, from defining the study scope and selecting product features to probabilistic generation, post-processing, and final export. This is the framework that integrates the logical rule generation, probabilistic modeling, and AI-assisted reasoning in order to develop a synthetic dataset for the laptop industry in the mass customization context. The generated dataset serves as a well-structured and scalable foundation for a future study of demand forecasting and procurement planning.

To improve the clarity and reproducibility of the dataset generation process, the overall procedure is summarized in the following algorithm. It specifies the probabilistic sampling steps and the constraints verification applied during the generation of each synthetic transaction:

Input:

1) Customer segments S
2) Customer attributes distributions $P(C|S)$
3) Product attribute distributions $P(A|S)$
4) Temporal distribution $P(T)$
5) Compatibility constraint set Ω
6) Number of transactions N

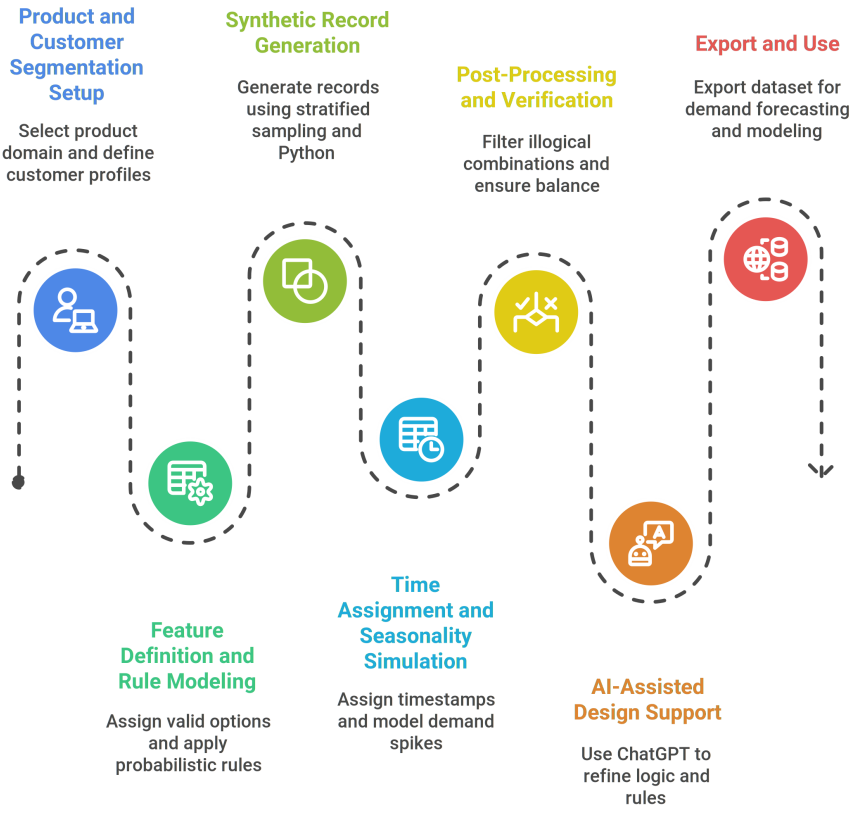
Output:

1) Synthetic dataset D

Steps:

1) Initialize empty dataset D
2) For i = 1 to N do:

Figure 4
Synthetic data generation process



a. Sample segment $S_i \sim P(S)$
b. Sample customer attributes $C_i \sim P(C|S_i)$
c. Sample product attributes $A_i \sim P(A|S_i)$
d. Sample timestamp $T_i \sim P(T)$
e. If $A_i \in \Omega$, accept the configuration and add $x_i = (S_i, C_i, A_i, T_i)$ to dataset D

3) Return dataset D

In the next section, we present the validation methods used to check the statistical realism and logical consistency of the generated dataset across customer segments and product configurations.

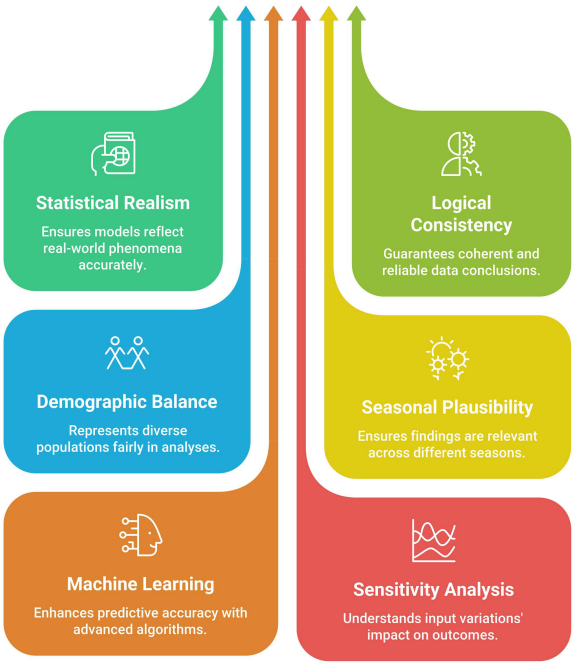
3.8. Synthetic dataset validation process

Given the absence of any open-access dataset that describes demand behavior in the mass customization context and the fact that the only datasets that exist are proprietary, commercially restricted, or not disclosed in published studies, direct comparison with real-world data is not possible. Consequently, the validation strategy combines statistical, logical, demographic, temporal, and predictive utility testing. And so, based on this combination, we could verify that the generated data respected realistic feature distributions and domain relations between products and customers with interpretability guarantees and methodological transparency.

To illustrate the validation logic used in this work, Figure 5 illustrates the six dimensions considered in the evaluation.

Statistical realism was used to check whether the generated variables followed plausible real-world distributions. Logical consistency examined whether the links between customer profiles and product attributes respected realistic domain rules. Demographic balance assessed the representation of different customer segments in the dataset. Seasonal plausibility was then verified whether demand over time reflected expected purchasing cycles and seasonal variations. A predictive validation step based on machine learning models is conducted to assess whether the generated dataset contains coherent and learnable structures that can support forecasting experiments. Finally, sensitivity analysis evaluates the robustness of the framework by examining how moderate variations in key assumptions affect the generated demand structure and predictive utility.

Figure 5
Dataset validation framework



3.9. Machine learning-based predictive validation

In addition to the statistical and logical validation, the study included a predictive validation step. This step examined whether the synthetic dataset contained coherent and learnable patterns that could support demand forecasting in a mass customization context.

Since no real-world dataset was available for direct comparison, this validation did not measure actual forecasting accuracy. Instead, it assessed the predictive utility of the generated data. The aim was to verify whether machine learning models could learn consistent relationships between customer profiles, product configurations, and temporal demand patterns.

Two forecasting angles were looked at together since each one captures something the other misses. The first focuses on component-level demand prediction, where weekly demand was aggregated for individual product attributes such as CPU, RAM, or GPU. Categorical variables were encoded, and a seasonal representation was introduced using sinusoidal transformations of the week index. A Light Gradient Boosting Machine (LightGBM) model was then trained to predict weekly component demand. This model was selected due to its ability to handle heterogeneous tabular data and capture nonlinear relationships between product and temporal demand patterns.

The second perspective evaluates temporal demand dynamics. A univariate weekly demand series was constructed by aggregating all laptop transactions across the synthetic year. A Long Short-Term Memory (LSTM) neural network was then trained to predict short-term changes in demand. The model used sliding windows of six consecutive weeks, which allowed it to learn from recent demand history before estimating the next period. This recurrent structure was selected because it can capture temporal dependencies and seasonal signals present in the generated dataset.

Both models were trained using an 80/20 split between training and testing data. Their performance was evaluated with standard forecasting metrics: mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). The results are presented in the next section. They provide an additional validation layer showing that the generated dataset supports machine learning-based forecasting experiments.

All machine learning experiments were implemented in Python using standard libraries including LightGBM and TensorFlow/Keras.

3.10. Sensitivity analysis design

A scenario-based sensitivity analysis was carried out to assess the robustness of the proposed framework under moderate changes in the main modeling assumptions. The dataset was generated from several predefined assumptions. These assumptions concerned the relative size of each customer segment, the probability of selecting product attributes within each segment, and the seasonal weights used to distribute demand over time. For this reason, a sensitivity analysis was added to examine how much these assumptions affected the final demand structure. The same analysis also checked whether changes in the assumptions changed the predictive usefulness of the dataset.

The analysis started from a baseline scenario, which corresponded to the original generated dataset. Three additional scenarios were then created by modifying one assumption at a time.

The first scenario focused on customer composition. The share of the student segment was increased by 10%, while the share of the gaming enthusiast segment was reduced by 10%. This made it possible to observe how a shift in the population structure affected the generated demand.

The second scenario focused on product preferences within one segment. For gaming enthusiasts, the probability of selecting high-performance GPUs was reduced by 10%. The probabilities of the other GPU options in the same segment were then redistributed proportionally, so the total probability remained equal to one. This scenario tested whether the configuration mix was sensitive to changes in conditional preferences.

The third scenario focused on seasonality. The weights assigned to the two main demand peaks, August–September and November–December, were reduced by 10%. The remaining monthly weights were then adjusted proportionally to keep the total demand volume unchanged. This scenario examined how the dataset reacted when the strength of the main seasonal peaks was weakened.

Each scenario was evaluated through descriptive indicators that captured changes in the generated demand structure. These indicators included customer segment shares; the proportion of integrated GPUs; the combined share of RTX 3060, RTX 3070, and RTX 3080 configurations; the share of demand observed in August–September; and the share observed in November–December. Predictive usefulness was also assessed using the same forecasting-oriented evaluation procedure applied during the validation stage. The purpose of this analysis was not to optimize the generation parameters but rather to verify whether moderate variations in the underlying assumptions produce substantial changes in the generated demand patterns or in the dataset's capacity to support learning.

4. Results

Following the validation framework described in Section 3, the dataset was evaluated along six main dimensions:

1) Statistical realism of individual features,
2) Logical consistency between customer segments and product choices,
3) Demographic balance across customer segments,
4) Seasonal plausibility of demand trends,
5) Machine learning-based predictive analysis,
6) Sensitivity analysis.

These dimensions reflect accepted techniques for validation of synthetic data and consider the specificities of configurable and make-to-order environments.

4.1. Dataset overview

The final dataset contains 500,000 synthetic laptop sales transactions in a complete annual cycle. Each transaction combines customer demographics attributes, product specifications, and a timestamp. The dataset was constructed to incorporate realistic differences in the product selection and the behaviors of customers who select the products, thus balancing each segment of the five predefined consumer profiles.

4.2. Statistical realism of individual features

The statistical distribution of key technical features reinforced the statistical realism of the dataset.

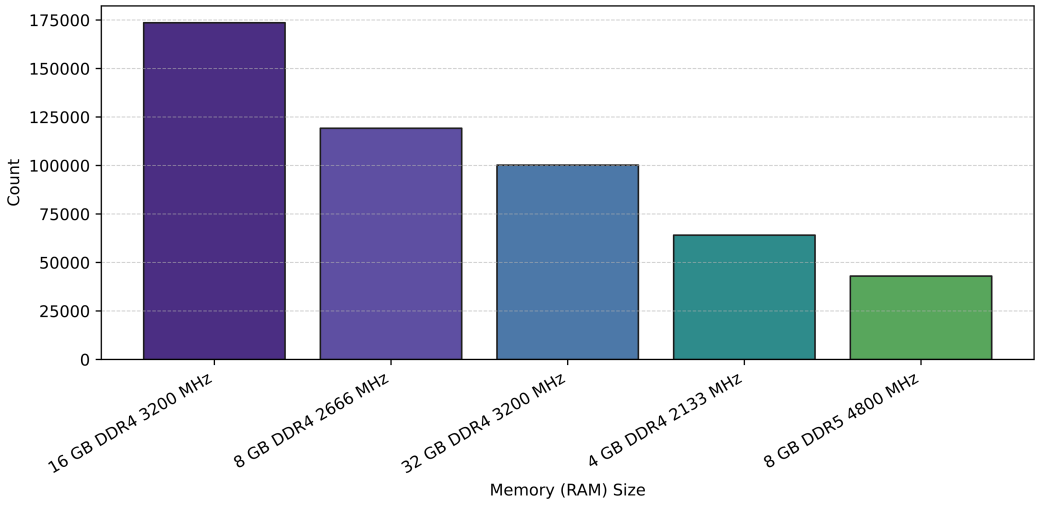
As shown in Figure 6, 16 GB DDR4 3200 MHz was the dominant RAM option, followed by 8 GB DDR4 2666 MHz and 32 GB DDR4 3200 MHz, while 4 GB DDR4 2133 MHz and 8 GB DDR5 4800 MHz appeared less frequently. This pattern fits a configurable laptop market where mid-range builds tend to dominate while higher-capacity options still hold a strong share thanks to gaming and professional users. A similar pattern appeared for GPU, storage, and screen size, with the higher-end choices concentrated in the gaming and professional segments.

These results demonstrate that probabilistic modeling, paired with AI-assisted reasoning, can reproduce realistic demand behavior in a mass customization setting without overstating the share of extreme values.

4.3. Logical consistency between customer segments and product choices

Beyond univariate distributions, logical consistency between customer profiles and selected configurations was assessed

Figure 6
Distribution of RAM sizes



through rule-based and visual verification. In mass customization environments, different customer segments tend to show distinct but nondeterministic preference patterns depending on their needs and usage contexts. A key part of assessing the dataset's realism is checking whether customer profiles align logically with selected laptop features.

Figure 7 focuses on RAM choices across the different customer profiles. The heatmap shows that the generated data does not assign the same memory capacity to all customers in a given segment. Instead, it produces clear but flexible tendencies. Gaming enthusiasts and tech professionals are more often associated with larger RAM capacities, which is consistent with their need for performance-oriented devices. Students appear more frequently in the lower and intermediate RAM categories, while general consumers are mostly centered around mid-range options. This distribution keeps the expected segment-specific behavior without making the configurations overly rigid.

Figure 8 examines the link between RAM capacity and storage type. The link between these two components is not fixed in the same way as customer-driven attributes are. RAM and storage

Figure 7
Heatmap between customer profile and RAM capacity

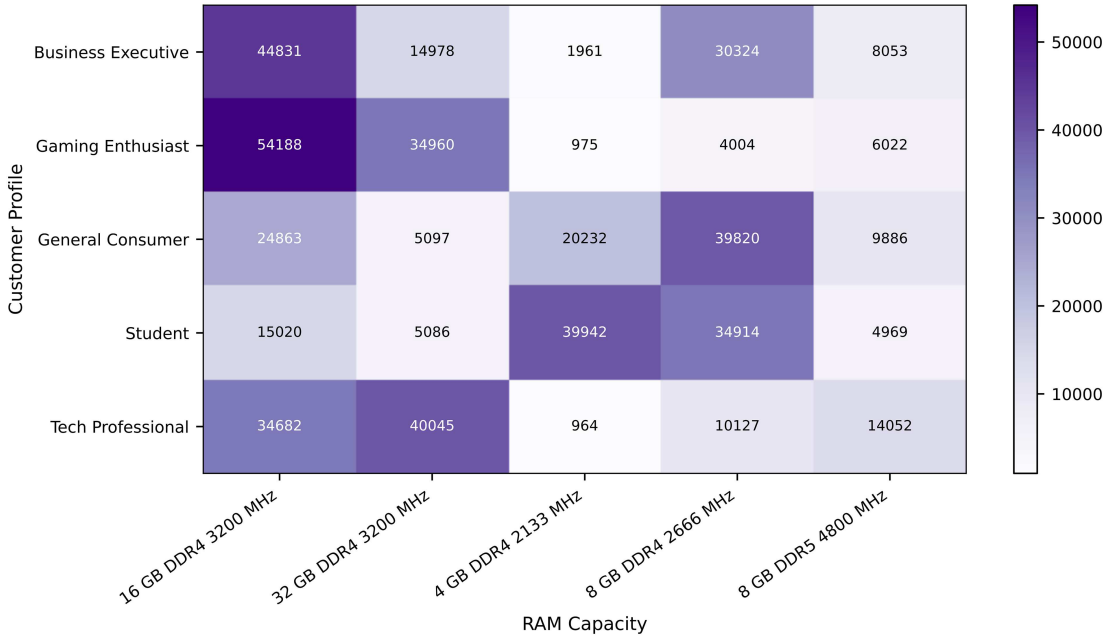
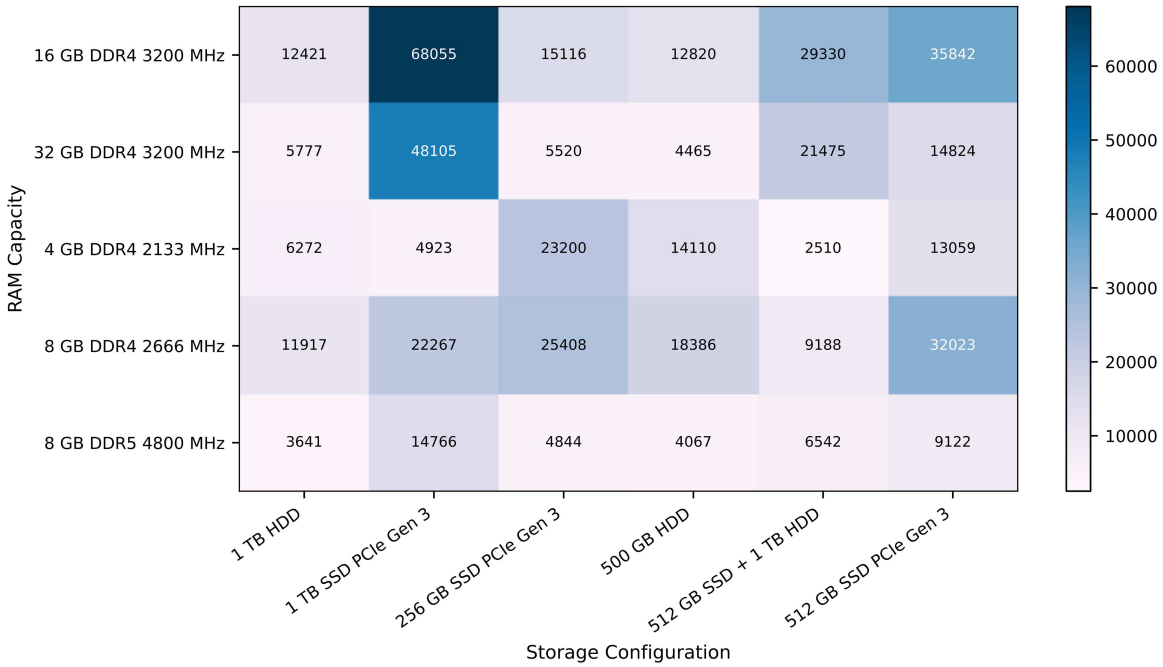


Figure 8
Heatmap between RAM capacity and storage configuration



can appear in different combinations because laptop configurations often depend on price range and market positioning. This makes some combinations more common than others, but it does not impose a fixed relationship between the two components. The observed distribution confirms that the dataset preserves configuration diversity without imposing deterministic relationships between product attributes.

Together, these analyses demonstrate that the probabilistic generation framework successfully captures both behavioral purchasing tendencies and realistic configuration flexibility, which are characteristic of mass customization product offerings.

4.4. Demographic balance across customer segments

From a demographic perspective, the dataset was equally distributed among the five target profiles: students, gamers, professionals, executives, and general consumers. Such a balance minimizes bias and allows the data to feed into global as well as segment-specific forecasting models.

4.5. Seasonal plausibility of demand trends

The temporal validation showed whether the synthetic demand reproduced a realistic seasonal purchasing cycle as it would be expected. The generated timeline showed familiar sales cycles, with spikes in August–September, which represent the back-to-school period; November–December, which is the holiday season; and March–April, which coincides with the professional upgrades. Such seasonal shifts align with known behavior in electronics market dynamics.

The aggregated monthly demand distribution is shown in Figure 9, thus proving that the synthetic dataset captures temporal dynamics instead of simply a uniform or random distribution of time.

4.6. Machine learning-based predictive analysis

To complete the statistical and logical validation analyses, two machine learning models were trained to evaluate whether the synthetic dataset contains coherent and learnable patterns relevant for demand forecasting tasks.

The LightGBM model achieved strong predictive performance as shown in Table 3, indicating that the synthetic dataset encodes coherent relationships between product configurations, customer attributes, and weekly demand levels. While the LSTM model reproduced the temporal demand dynamics embedded in the dataset with reasonable accuracy, demonstrating the coherence of the seasonal structure generated during the simulation process.

In sum, the validation results show the proposed dataset to exhibit sufficient statistical fidelity, behavioral logic, demographic balance, and temporal plausibility, and predictive utility. This

Figure 9
Overall annual pattern of laptop sales

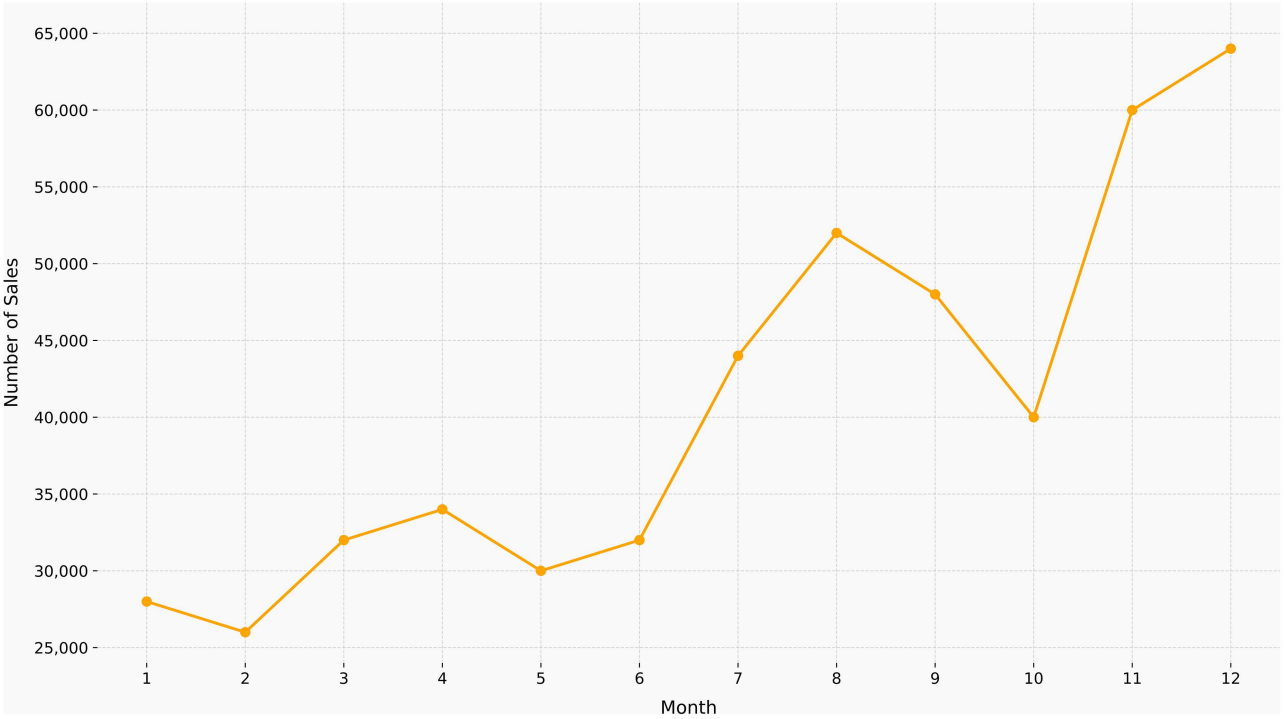


Table 3
Predictive validation results of machine learning models

Model	Target	MAE	RMSE	R^2
LightGBM	Weekly component demand	161.30	217.33	0.9293
LSTM	Weekly total demand	1.75	2.92	0.6159

confirms its suitability for demand forecasting research in mass customization environments.

4.7. Sensitivity analysis results

The scenario-based sensitivity analysis assessed the effect of moderate perturbations applied to three main groups of assumptions, namely, customer segment composition, gaming-oriented GPU preference weights, and seasonal peak intensities. The main results obtained for the baseline and perturbed scenarios are reported in Table 4.

In the baseline scenario, the student and gaming enthusiast segments accounted for 19.99% and 20.03% of the dataset, respectively. The share of configurations using integrated GPUs reached 27.16%, whereas RTX 3060/3070/3080 configurations represented 29.24%. From a temporal perspective, August–September represented 20.41% of total demand, whereas November–December accounted for 25.31%. The predictive utility proxy produced an MAE of 1370.79, an RMSE of 2053.04, and an R^2 value of 0.8188.

When customer segment composition was perturbed, the student share increased to 21.99%, while the gaming enthusiast share declined to 18.03%. This change was accompanied by an increase in the integrated GPU share to 28.31% and a decrease in the share of RTX 3060/3070/3080 configurations to 28.31%. Despite these variations, predictive utility remained very close to the baseline, with an R^2 value of 0.8173.

A similarly limited effect was observed when the gaming-oriented GPU preference weights were modified. In this case, the integrated GPU share rose to 27.69%, while the share of RTX 3060/3070/3080 configurations declined to 28.55%. The overall segment structure remained close to the baseline, and predictive utility also remained stable, with a 0.8178.

The most visible effect was observed when the seasonal peak intensities were reduced. Under this perturbation, the demand share of August–September decreased from 20.41% to 18.51%, while that of November–December fell from 25.31% to 22.95%. In this scenario, the predictive utility proxy showed the largest variation among the tested cases, with the R^2 value increasing to 0.8764.

Overall, the perturbations applied to customer segment composition and conditional preference weights produced only limited changes in the generated demand structure and in predictive utility. By contrast, the reduction of seasonal peak intensities produced the strongest variation among the tested scenarios, indicating that temporal weighting is the most influential assumption group within the current framework.

5. Discussion

The goal of this study is to generate and validate a synthetic dataset that can reflect demand behavior in a mass customization context, where high variety and lack of access to real data are challenging the classic forecasting researches. The findings in the previous section confirm that the aim set above is achieved in various complementary aspects.

From a statistical perspective, the distributions of the characteristics that are evident from the dataset ensure that the realistic market behaviors are well represented without concentration on extreme configurations. Dominant component choices such as mid-range RAM capacities and balanced storage options reflect common purchasing behavior, while higher-end configurations remain present but limited to specific customer profiles. This balance is essential for demand forecasting applications, as it prevents bias toward either low-end or high-end products while preserving sufficient variability for model training.

One more major strength of the proposed dataset, aside from the univariate realism, is the coherent but nondeterministic relationships between customer profiles and configuration choices. The generated data does not treat customer profiles and laptop configurations as unrelated variables. Instead, it keeps the behavioral tendencies that link the two, while still leaving room for variation inside each segment. Gaming enthusiasts and tech professionals appear more often with larger RAM capacities, which matches their stronger need for performance. Students, on the other hand, are more frequently associated with lower and intermediate memory options. This matters for mass customization because small differences in customer behavior can change the expected demand for components and affect procurement planning.

The link between RAM and storage gives another indication of internal coherence. The dataset does not force a fixed combination between these two attributes. It allows several technically plausible configurations, while making some combinations more common than others. This balance is important because realistic laptop markets include repeated patterns, but they also include diversity across product ranges and customer needs.

Temporal validation adds a further layer to this realism assessment. The fact that there are peaks of demand corresponding to the purchasing cycles in the electronics market indicates

Table 4
Results of the scenario-based sensitivity analysis

Scenario	Student share (%)	Gaming enthusiasts share (%)	Integrated GPU share (%)	RTX 3060 /3070 /3080 share (%)	August–September share (%)	November–December share (%)	MAE	RMSE	R^2
Baseline	19.99	20.03	27.16	29.24	20.41	25.31	1370.79	2053.04	0.8188
Segment distribution shift	21.99	18.03	28.31	28.31	20.41	25.31	1367.12	2034.22	0.8173
Gaming GPU preference shift	20.33	18.67	27.69	28.55	20.41	25.31	1370.75	2048.25	0.8178
Seasonal peak reduction	19.99	20.03	27.17	29.23	18.51	22.95	1061.28	1563.80	0.8764

that the generation process has captured temporal dynamics. This has immense importance in research related to forecasting since it becomes possible to assess different models on the basis of realities of demand variation rather than a fixed demand.

The results of the predictive validation experiments also confirm these results. The LightGBM model reached high explanatory power in predicting the component-level demand, which means that the relationships between customer profiles, product configurations, and demand levels are coherent and learnable.

The temporal results add another important check to the validation process. The LSTM model captured the main structure of demand over time and generated stable short-term forecasts. This result suggests that the dataset does not contain only random variation. It includes regularities that forecasting models can detect and use.

This predictive test should be interpreted carefully. The objective was not to prove that the model would achieve the same accuracy in an industrial setting. Such a comparison is not possible here, since no open reference dataset exists for demand in mass customization contexts. The goal was more specific to examine whether the synthetic data contains coherent and learnable patterns that can be used in forecasting experiments.

The proposed dataset also contributes to ongoing work on synthetic data generation and demand modeling. Many earlier studies use aggregated sales records, simplified product structures, or confidential industrial datasets. These sources are useful, but they often limit reproducibility. The dataset developed in this study takes a different direction. What makes it stand apart is that it brings together customer information, configuration attributes, and temporal demand behavior in one reproducible and solid structure. For this reason, it should be viewed as a complementary resource rather than a substitute for real industrial data. It also responds to a gap that recent work on mass customization and forecasting has pointed out.

From a practical point of view, the dataset can support several types of analysis. Researchers can use it to compare forecasting models, study demand disaggregation, or test procurement-oriented decision support methods. It also allows experiments in which customer preferences vary jointly with product configurations, which is important for realistic modeling in mass customization. Forecasting models such as Random Forest, Gradient Boosting, and recurrent neural networks can be applied at both the configuration level and the component level, for example, for CPU or GPU demand. The same structure can also support procurement planning through component demand estimation, safety stock analysis, and supplier capacity planning in high-variety make-to-order environments.

The dataset has already proven useful in practice. A recent study by El Assad et al. [30] relied on it as the experimental basis for the work that followed. In that work, it was used as the experimental basis for a multilevel planning approach in Configure-to-Order/Assemble-to-Order (CTO/ATO) mass customization supply chains. The dataset made it possible to transform demand patterns into planning decisions and to evaluate improvements in lead-time performance. This application shows that the dataset is not limited to statistical validation. It can also support downstream decision support and optimization tasks in supply chain contexts.

The synthetic nature of the dataset does not mean that its construction was arbitrary. Customer segments, configuration options, and probability distributions were defined using domain knowledge and literature on laptop markets, consumer behavior, and configurable products. The probabilistic framework was built to maintain realistic relationships between variables, while still allowing variation and occasional deviations. This is important because real demand does not follow perfectly fixed rules. The generation rules and constraints were also defined explicitly and applied consistently, which strengthens the transparency and reproducibility of the process.

However, the generation process remains based on assumptions. These assumptions mainly concern customer segmentation, conditional attribute distributions, and the stability of relationships over time.

These assumptions shape the resulting demand patterns, and variations in the associated probability weights may influence forecasting outcomes. However, the objective of this study is not to replicate a specific real-world dataset but to provide a controlled and adaptable environment for evaluating forecasting models. The generation framework was not designed as a fixed model. Its assumptions can be revised, and its parameters can be recalibrated when researchers apply it to another product family, market, or industrial context.

The sensitivity analysis gives a clearer view of this adaptability. When customer segment shares and conditional preference weights were moderately changed, the dataset did not lose its internal consistency. These changes mainly shifted the relative presence of certain configurations. They did not fundamentally alter the overall structure of the data or its usefulness for forecasting experiments. The largest variation appeared when the seasonal weights were modified. This is expected, since seasonality directly shapes the distribution of demand over time and affects the metrics used in forecasting evaluation. This result suggests that the framework is generally stable, but that seasonal assumptions should be adjusted carefully in future applications.

Despite these strengths, the dataset remains a synthetic representation of demand. It cannot reproduce all the factors that influence real markets. For example, it does not include sudden disruptions, price changes, promotional campaigns, competitor behavior, or hidden customer motivations. The assumptions used in the model are grounded in domain knowledge and literature, but they should be interpreted as context-specific choices rather than universal rules.

Future work can build on this structure in several directions. The dataset could be adapted to other customizable products or extended to represent changes in customer preferences over time. It could also include external variables, such as price variations or macroeconomic indicators. These additions would make the dataset more suitable for broader forecasting studies and for procurement-oriented optimization in different industrial settings.

6. Conclusion

This study developed a synthetic dataset to support demand forecasting research in mass customization, where real open-access data remain limited.

The proposed dataset uses the laptop industry as its application case. This choice is relevant because laptops combine modular components, customer-driven configuration choices, and strong variation in demand. The dataset included 500,000 simulated transactions across one year. Each transaction links customer characteristics, configurable product attributes, and seasonal purchasing behavior. This structure was designed to reflect the type of complexity that forecasting models face in high-variety product environments.

The dataset was not evaluated only by checking its size or format. Its realism was examined through several validation steps. These included the analysis of segment balance, the verification of logical relationships between customer profiles and configurations, and the study of statistical distributions. These checks looked at the dataset from three angles: how well it fits the rules that shape the laptop market, how its distributions hold up under inspection, and how ready it is for use in machine learning models.

The results show that the dataset reflects real-world trends, stays consistent with the logic of the field, and shows statistical properties that look realistic. It is also built for use in data-driven forecasting work carried out on real problems. For researchers and practitioners, it offers a strong starting point for building, evaluating, and comparing forecasting models in mass customization settings. This is important because traditional planning methods often struggle when demand is fragmented across many product variants and component combinations.

Future work can build on this dataset in several directions, including multilevel forecasting strategies. Further developments may also introduce changing customer preferences, behavioral feedback such as satisfaction or repeat purchases, and comparisons with real sales data when such data become available. Through these extensions, the dataset will contribute to advancing demand forecasting, procurement planning, and supply chain decision-making in a configurable product environment shaped by Industry 4.0 principles.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://doi.org/10.34740/kaggle/dsv/15258985.

Author Contribution Statement

Nouhaila El Assad: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Visualization. **Kawtar El Haouti:** Methodology, Validation. **Soumaya Zayrit:** Software, Data curation. **Salah-eddine Mokhlis:** Validation, Formal analysis. **Mohamed Rhouzali:** Formal analysis, Investigation. **Najat Messaoudi:** Validation, Writing – review & editing, Supervision, Project administration.

References

[1] Hu, S. J. (2013). Evolving paradigms of manufacturing: From mass production to mass customization and personalization. *Procedia CIRP*, *7*, 3–8. https://doi.org/10.1016/j.procir.2013.05.002

[2] Wang, Y., Ma, H.-S., Yang, J.-H., & Wang, K.-S. (2017). Industry 4.0: A way from mass customization to mass personalization production. *Advances in Manufacturing*, *5*(4), 311–320. https://doi.org/10.1007/s40436-017-0204-7

[3] Jain, P., Garg, S., & Kansal, G. (2023). A TISM approach for the analysis of enablers in implementing mass customization in Indian manufacturing units. *Production Planning & Control*, *34*(2), 173–188. https://doi.org/10.1080/09537287.2021.1900616

[4] Jamouli, Y., Tetouani, S., Cherkaoui, O., & Soulhi, A. (2023). A model for Industry 4.0 readiness in manufacturing industries. *Data and Metadata*, *2*, 200. https://doi.org/10.56294/dm2023200

[5] Jamouli, Y., Tetouani, S., Cherkaoui, O., & Soulhi, A. (2023). To diagnose Industry 4.0 by maturity model: The case of Moroccan clothing industry. *Data and Metadata*, *2*, 137. https://doi.org/10.56294/dm2023137

[6] El Assad, N., Dachry, A., Fouraiji, H., Dachry, W., & Messaoudi, N. (2025). Smart manufacturing paradigms in the context of Industry 4.0: Bibliometric analysis. *E3S Web of Conferences*, *601*, 00035. https://doi.org/10.1051/e3sconf/202560100035

[7] Kim, S., & Lee, K. (2023). The paradigm shift of mass customisation research. *International Journal of Production Research*, *61*(10), 3350–3376. https://doi.org/10.1080/00207543.2022.2081629

[8] Roy, M.-A., & Abdul-Nour, G. (2024). Integrating modular design concepts for enhanced efficiency in digital and sustainable manufacturing: A literature review. *Applied Sciences*, *14*(11), 4539. https://doi.org/10.3390/app14114539

[9] Christopher, M. (2016). *Logistics and supply chain management* (5th ed.). UK: Pearson.

[10] Terrada, L., Khaili, M. E., & Ouajji, H. (2022). Demand forecasting model using deep learning methods for Supply Chain Management 4.0. *International Journal of Advanced Computer Science and Applications*, *13*(5), 704–711. https://doi.org/10.14569/IJACSA.2022.0130581

[11] Yağcıoğlu, E., Tekin, A. T., & Çebi, F. (2022). Demand forecasting of a company that produces by mass customization with machine learning. In *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference*, 197–204. https://doi.org/10.1007/978-3-030-85577-2_23

[12] Kim, M., Jeong, J., & Bae, S. (2019). Demand forecasting based on machine learning for mass customization in smart manufacturing. In *Proceedings of the 2019 International Conference on Data Mining and Machine Learning*, 6–11. https://doi.org/10.1145/3335656.3335658

[13] Buggineni, V., Chen, C., & Camelio, J. (2024). Enhancing manufacturing operations with synthetic data: A systematic framework for data generation, accuracy, and utility. *Frontiers in Manufacturing Technology*, *4*, 1320166. https://doi.org/10.3389/fmtec.2024.1320166

[14] Liu, S.-F., Wang, S.-Y., & Tung, H.-H. (2024). A comprehensive decision-making approach for strategic product module planning in mass customization. *Mathematics*, *12*(11), 1745. https://doi.org/10.3390/math12111745

[15] Bortolini, M., Calabrese, F., Galizia, F. G., & Regattieri, A. (2023). A two-step methodology for product platform design and assessment in high-variety manufacturing. *The International Journal of Advanced Manufacturing Technology*, *126*(9), 3923–3948. https://doi.org/10.1007/s00170-023-11347-8

[16] Modrak, V., & Soltysova, Z. (2023). Assessment of product variety complexity. *Entropy*, *25*(1), 119. https://doi.org/10.3390/e25010119

[17] Flyckt, J., & Lavesson, N. (2025). Navigating demand forecasting in make-to-order manufacturing: The role of global models and intermittent time-series. *CEUR Workshop Proceedings, 4037*, 12–25.

[18] Hof, L. A., & Wüthrich, R. (2018). Industry 4.0 – Towards fabrication of mass-personalized parts on glass by Spark Assisted Chemical Engraving (SACE). *Manufacturing Letters*, *15*, 76–80. https://doi.org/10.1016/j.mfglet.2017.12.003

[19] Dai, S. (2023). Short-term load forecasting with deep learning techniques. *Journal of Physics: Conference Series*, *2547*(1), 012025. https://doi.org/10.1088/1742-6596/2547/1/012025

[20] Neira Villar, J. R., & Cano Lengua, M. A. (2024). A systematic review of the literature on the use of artificial intelligence in forecasting the demand for products and services in various sectors. *International Journal of Advanced Computer Science and Applications*, *15*(3), 144–156. https://doi.org/10.14569/IJACSA.2024.0150315

[21] Chan, K. C., Rabaev, M., & Pratama, H. (2022). Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation. *Production & Manufacturing Research*, *10*(1), 337–353. https://doi.org/10.1080/21693277.2022.2086642

[22] Yadav, P., Gaur, M., Fatima, N., & Sarwar, S. (2023). Qualitative and quantitative evaluation of multivariate time-series synthetic data generated using MTS-TGAN: A novel approach. *Applied Sciences*, *13*(7), 4136. https://doi.org/10.3390/app13074136

[23] Yoo, H., Moon, J., Kim, J.-H., & Joo, H. J. (2023). Design and technical validation to generate a synthetic 12-lead electrocardiogram dataset to promote artificial intelligence research. *Health Information Science and Systems*, *11*(1), 41. https://doi.org/10.1007/s13755-023-00241-y

[24] Raizada, S., & Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting. *International Journal of Advanced Computer Science and Applications*, *12*(11), 102–110. https://doi.org/10.14569/IJACSA.2021.0121112

[25] Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, 513–520. https://doi.org/10.7551/mitpress/7503.003.0069

[26] Pachouly, J., Ahirrao, S., Kotecha, K., Selvachandran, G., & Abraham, A. (2022). A systematic literature review on software defect prediction using artificial intelligence: Datasets, data validation methods, approaches, and tools. *Engineering Applications of Artificial Intelligence*, *111*, 104773. https://doi.org/10.1016/j.engappai.2022.104773

[27] Lázaro, C., & Angulo, C. (2024). Using UMAP for partially synthetic healthcare tabular data generation and validation. *Sensors*, *24*(23), 7843. https://doi.org/10.3390/s24237843

[28] El Assad, N., Mokhlis, S.-E., & Messaoudi, N. (2026). Generation and validation of a synthetic dataset for demand modelling in mass customization using artificial intelligence tools. In *Connected Objects, Artificial Intelligence, Telecommunications and Electronics Engineering: Proceedings of COCIA'2025*, 529–535. https://doi.org/10.1007/978-3-032-01536-5_80

[29] Shi, J., Huang, F., Jia, F., Yang, Z., & Rui, M. (2023). Mass customization: The role of consumer preference measurement, manufacturing flexibility and customer participation. *Asia Pacific Journal of Marketing and Logistics*, *35*(6), 1366–1382. https://doi.org/10.1108/APJML-10-2021-0719

[30] El Assad, N., Mokhlis, S., El Haouti, K., & Messaoudi, N. (2026). Development of a lead-time-first multi-level planning approach for CTO/ATO mass customization supply chains. *Eastern-European Journal of Enterprise Technologies*, *1*(3(139)), 48–60. https://doi.org/10.15587/1729-4061.2026.345253

How to Cite: Assad, N. E., Haouti, K. E., Zayrit, S., Mokhlis, S.-eddine, Rhouzali, M., & Messaoudi, N. (2026). AI-Assisted Synthetic Dataset Generation and Validation for Demand Forecasting in the Context of Mass Customization. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62029121