ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

School of Computer Science and Engineering
Master Degree in Artificial Intelligence

UNDERSTANDING AND PREDICTING CUSTOMER CHOICES THROUGH ADVANCED MACHINE LEARNING FORECASTING: THE DUCATI MOTORCYCLES CASE STUDY

Thesis in
Machine Learning

Supervisor
Chiar.mo Prof.Ing. Claudio Sartori

Co-supervisor
Dott. Massimiliano Bertei

Candidate
Irene Burri

Third Session
Academic Year 2025-2026

Keywords

Demand Forecasting
Supply Chain Planning
Automotive Industry
Machine Learning Forecasting
Predictive Analytics

A Spillo

Table of contents

List of Tables

List of Figures

Chapter 1

Introduction

The increasing complexity of modern manufacturing and the growing demand for product personalization are reshaping the way companies plan, forecast and manage their operations. As production systems become more flexible and product portfolios more configurable, traditional forecasting approaches struggle to capture the variability and interdependence of customer choices. In this context, artificial intelligence and machine learning are emerging as key enablers for data-driven decision-making, offering manufacturers the ability to model nonlinear patterns, incorporate behavioural signals and anticipate demand with greater precision, and, crucially, to do so in an automated and scalable manner.

The automotive and motorcycle industries are undergoing this transformation as well. Optional components have evolved from simple accessories into strategic elements of product differentiation, directly influencing customer satisfaction, production efficiency, and supply chain performance. As manufacturers progressively integrate configurability into the purchasing process, the ability to anticipate customer choices becomes essential to ensure production continuity and avoid costly imbalances between supply and demand.

Ducati Motor Holding is undergoing this transformation. With the recent introduction of configurability at the time of purchase, optional components have become integral to the assembly workflow, making the ability to anticipate customer choices essential for production continuity. This shift introduces new challenges for demand planning, particularly when suppliers operate with long lead times.

In this context, forecasting the take rate of optional components, defined as the proportion of motorcycles expected to include a given optional, becomes a key capability for planning and procurement teams.

Accurate take rate predictions support inventory optimization, reduce the

risk of production delays and improve alignment between sell-in and sell-out volumes.

The objective of this thesis is to design and evaluate a machine learning forecasting framework capable of predicting the monthly take rate of optional components for configurable motorcycle models, using a real industrial case study to ground the analysis.

The forecasting pipeline integrates multiple methodological components: order level binary classification, monthly aggregation, synthetic temporal augmentation to extend the limited historical horizon and association rule-based features to capture co-occurrence patterns among optional components.

The main contributions of this thesis can be summarised along four complementary directions. First, it proposes a machine learning forecasting architecture specifically designed to accommodate Ducati's operational constraints and the structure of its available data. Building on this foundation, the work develops a modelling strategy capable of sustaining a long prediction horizon, such as twelve months, despite the limited historical depth of the dataset. To address this challenge, the pipeline incorporates synthetic data generation techniques that help reconstruct temporal patterns and enhance long term forecasting stability. Finally, the thesis introduces association rule-based features, enriching the predictive space with behavioural signals derived from customer configuration choices and capturing interactions that would otherwise remain latent.

The thesis is structured as follows. Chapter 2 provides the industrial and organizational context, describing Ducati's background and the role of optional components in production. Chapter 3 presents the theoretical background, including forecasting concepts, take rate modelling and relevant machine learning techniques. Chapter 4 details the development of the forecasting pipeline, from data collection to model design. Chapter 5 presents the experimental results and evaluates the performance of the proposed models. Finally, Chapter 6 concludes the thesis and outlines potential directions for future work.

Chapter 2

Context and General Background

The automotive industry is undergoing a significant transformation, driven by increasing product complexity, evolving consumer expectations and heightened uncertainty across global markets. Recent literature highlights how socio-economic conditions, technological developments and external environmental factors contribute to highly variable demand patterns, making forecasting more challenging and strategically relevant than in the past [1, 2]. At the same time, the sector faces recurrent supply chain disruptions caused by geopolitical tensions, economic fluctuations and unpredictable events, which expose structural vulnerabilities and reinforce the need for more resilient and data-driven planning processes [3].

From the consumer perspective, preferences are becoming increasingly heterogeneous. According to Deloitte's Global Automotive Consumer Study, customers now prioritize personalization, digital features and advanced technologies, with configuration choices playing a growing role in purchase decisions [4]. This shift increases the variability of demand forecasting approaches, which were historically designed for more stable and homogeneous product portfolios.

In this context, demand forecasting has therefore become a strategic capability for automotive manufacturers. A recent systematic review shows that forecasting models support production planning, inventory management, procurement and long-term strategic decisions, particularly in industries characterized by high product complexity and volatile demand [1].

Accurate forecasts enable companies to optimize capacity, reduce inventory costs, mitigate stockout risks and improve overall operational efficiency.

A particularly relevant trend is the growing importance of personalization. Studies indicate that consumers increasingly value configurable products and tailored services, pushing manufacturers to expand modular offerings and

integrate personalization into their commercial strategies [5, 6].

This evolution aligns with broader shifts described in the mass customization literature, where customers are no longer passive recipients but become "active players in the value chain", positioned at the center of value creation [7].

As a result, forecasting no longer concerns only aggregate vehicle volumes but also the composition of demand, including the selection of optional components.

To ground these considerations in a concrete industrial setting, this study examines Ducati Motor Holding (DMH), an Italian manufacturer of high-performance motorcycles, characterized by a strong racing heritage, a consolidated design tradition and a technologically advanced product portfolio.

Ducati's global distribution network, combined with rising product configurability, makes precise forecasting of optional components vital for production continuity and service levels.

This reflects a broader phenomenon: as Piller and Tseng note, the shift toward individualized products introduces new operational challenges and requires firms to develop processes capable of managing variability, customer interaction and synchronization across the value chain [7].

The case study examined reflects the broader dynamics outlined above. Historically, optional components were installed after the vehicle had been produced allowing customers to add accessories independently and, most importantly, enabling manufacturers to manage inventories with considerable flexibility letting the companies decouple optional demand from production constraints. Customers or dealers purchased accessories independently and their availability was not directly linked to the assembly process.

Under this previous model, optional components did not influence the production workflow, since they were added only after the motorcycle delivery. As a result, forecasting accuracy was helpful but not critical for ensuring smooth operations.

The introduction of configurability at the time of purchase changes this paradigm. For some bike models included in this study, optional components are now selected during the ordering process increasing the relevance of personalization within the production workflow.

This shift reflects a set of implications commonly encountered in firms operating with configurable product architectures:

- Production dependency: production continuity depends on the timely availability of pre-selected components.
- Inventory pressure: optional components must be procured in advance, increasing the risk of stockouts or excess inventory. Stock shortages

may lead to production delays, extended lead times and potential loss of sales, while excessive inventory generates unnecessary holding costs and reduces warehouse efficiency, limiting the space available for other critical components.

- Forecasting complexity: optional demand becomes more volatile and model specific, requiring more granular and adaptive forecasting methods.

As a result, the ability to anticipate optional demand with high precision has become more important to ensure production continuity, avoid bottlenecks and maintain an optimal balance between service level and inventory cost.

The challenge is amplified by the limited historical data available. Configurability was introduced recently and each bike model has a different launch date. As a result, the time series of optional selections is short, irregular and influenced by structural events such as product rollouts, which generate artificial peaks in the order flow.

Accurate forecasting of optional take rates is therefore essential for:

- ensuring that all required components are available when needed;
- reducing emergency procurement and minimizing holding costs;
- avoiding unnecessary stock accumulation in warehouses;
- preventing production delays caused by missing optional components.

Given these challenges, Ducati requires a machine learning forecasting solution that is robust to limited historical depth, capable of capturing behavioural patterns in customer choices and flexible enough to support a long planning horizon.

Chapter 3

Theoretical Background

This chapter presents the theoretical foundations of the forecasting framework developed in this thesis. It introduces key concepts in demand forecasting, defines the notion of take rate, reviews machine learning approaches relevant to forecasting tasks and describes the methodological tools employed in the project. The aim is to provide a coherent background that motivates the modelling choices adopted in the subsequent chapters.

3.1 Demand Forecasting in Manufacturing

Demand forecasting refers to the set of quantitative methods used to estimate future customer demand based on historical observations, contextual variables and statistical or machine learning models.

Accurate demand forecasts play a central role in organizational planning, supporting activities such as procurement, scheduling, inventory control and long-term capacity decisions [8]. This need becomes even more pronounced in manufacturing environments, where operational efficiency and production continuity depend on the timely availability of components and resources.

Forecasting in industrial settings is often challenged by:

- seasonality and trend, which must be detected and modelled appropriately;
- irregular or intermittent demand, common in optional components or low volume products;
- short or incomplete historical series, especially for newly introduced products;

- structural breaks, such as product launches, regulatory changes or supply chain disruptions.

Classical time-series models such as ARIMA, exponential smoothing and state-space models assume relatively stable temporal patterns and sufficient historical depth [9]. When data is sparse, highly variable or influenced by structural events, these assumptions may not hold.

Machine learning approaches, by contrast, can incorporate heterogeneous features, capture nonlinear relationships and leverage cross-sectional information, making them suitable for complex forecasting tasks in modern manufacturing environments [10].

3.2 Take Rate: Definition and Importance

In the automotive industry, the take rate of an optional component quantifies the proportion of vehicles sold within a given period that support and include that option. Formally, for optional component k in period t:

$$\mathrm{TR}_{k,t} = \frac{N_{k,t}}{N_t} \tag{3.1}$$

where $N_{k,t}$ denotes the number of motorcycles sold in period t that support optional k and are configured with it, and N_t the total number of motorcycles sold in the same period that support optional k. The resulting value is a ratio between 0 and 1.

Take rate is widely used because it:

- normalizes demand across different production volumes;
- enables comparison across models, markets and time periods;
- supports long term planning independently of absolute sales forecasts;
- provides a stable metric for procurement and capacity decisions.

With high configurability, take rate forecasting becomes essential for ensuring component availability and avoiding production delays.

3.3 Machine Learning for Demand Forecasting

Demand forecasting can be approached through different modelling paradigms, which differ in how they represent temporal dynamics and how they exploit

available information. A first class of methods consists of traditional time series models, such as ARIMA, exponential smoothing, or state-space models, which directly model the temporal evolution of an aggregate demand indicator. These approaches are effective when long, stable and regularly spaced historical series are available, and when demand patterns can be captured through trend, seasonality and autocorrelation structures.

However, in many real-world settings, particularly those involving configurable products or rapidly evolving assortments, historical series may be short, irregular or disrupted by frequent product updates. In such cases, an alternative modelling strategy becomes advantageous, such as predicting demand at the level of individual transactions or choices and aggregating these predictions.

Here, the forecasting task is reframed as a supervised learning problem, where the model estimates the probability of a specific choice or event for each transaction, an approach widely explored in the demand forecasting literature [11]. Aggregate demand forecasts are then obtained by summing or averaging these individual probabilities. This formulation is theoretically grounded in the linearity of expectation, which ensures that aggregated predicted probabilities yield coherent estimates of overall demand.

This modelling strategy is particularly advantageous in situations where historical time series are short, irregular or frequently disrupted by product updates, which limits the effectiveness of traditional time-series approaches. It also becomes especially useful when each transaction is described by a rich set of features, allowing the model to exploit cross-sectional variation that would otherwise be lost in aggregate data. Moreover, when customer behaviour is characterised by nonlinear patterns or interaction effects, supervised learning methods offer the flexibility needed to capture these relationships, overcoming the constraints of classical statistical models.

Several machine learning models are well suited to this task. Tree-based methods, such as Random Forest and Gradient Boosting, are particularly effective because they can capture complex nonlinear relationships and high-order interactions without requiring explicit feature engineering. Their ability to handle heterogeneous data types and to remain robust with respect to outliers makes them well suited for modelling customer choice behaviour [12, 13].

Support Vector Machines offer a different perspective, in fact by maximising the margin between classes, they achieve strong generalisation performance, especially in high-dimensional settings. Through the use of kernel functions, SVMs can implicitly map the data into richer feature spaces, enabling the model to learn nonlinear decision boundaries without increasing the risk of

overfitting [14].

Neural networks provide an even more flexible modelling framework. Thanks to their layered structure, they can approximate highly complex functional relationships and capture subtle dependencies between features. Their expressiveness makes them suitable for scenarios where customer behaviour is influenced by multiple interacting factors, although they typically require larger datasets and careful regularisation.

Finally, Logistic Regression serves as a linear and interpretable baseline. Despite its simplicity, it provides well-calibrated probability estimates and allows for a clear understanding of how each feature contributes to the predicted outcome. Its limitations in modelling nonlinearities make it less competitive than more expressive models, but it remains a valuable benchmark for assessing the added value of more complex approaches.

This theoretical framework underpins many industrial applications where demand depends on individual configuration choices. The specific case analysed in this thesis, forecasting the selection of optional components in a configurable product, fits naturally within this modelling strategy.

3.4 Synthetic Data Generation

Synthetic data generation is increasingly adopted in industrial domains, such as manufacturing, automotive and supply-chain forecasting, where historical records are often limited, irregular or affected by structural changes. When the available dataset does not provide enough temporal depth or variability, artificially generated samples can enrich the training set while preserving the underlying statistical structure of the real data [15].

In these contexts, synthetic data helps reconstruct patterns that would otherwise be underrepresented, including rare configurations of features, long horizon temporal behaviours or combinations of attributes that occur infrequently in practice. This is particularly valuable for forecasting tasks, where model robustness depends on exposure to a sufficiently diverse set of historical scenarios.

Several methodological families can be used to generate synthetic observations, each capturing different aspects of the data-generating process:

- Probabilistic models: methods that estimate the joint distribution of the variables and sample from it (e.g. copula-based models, Bayesian networks, parametric multivariate distributions). Copula-based approaches have proven effective in modelling complex dependencies in tabular data [15].

- Generative machine learning models: approaches such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs) or diffusion models, which learn to reproduce complex, high-dimensional patterns.

- Resampling and bootstrapping techniques: strategies that create new sequences or records by recombining existing observations while maintaining temporal or structural coherence.

- Simulation-based approaches: domain-driven models that generate synthetic scenarios based on mechanistic rules, constraints or expert knowledge.

Across these approaches, the goal is to produce artificial samples that are statistically plausible and informative for the learning algorithm, without simply replicating the original data.

Frameworks such as the Synthetic Data Vault (SDV) exemplify this objective by providing modular generative models for realistic yet privacy-preserving synthetic datasets [16].

In the present work, synthetic data is employed to extend the effective training horizon and mitigate the limitations imposed by the short historical availability, improving the stability and generalisation of the forecasting models.

3.5 Association Rules

Association rule mining is widely used in industrial and commercial settings to uncover co-occurrence patterns among elements that appear together within configurations, processes or user choices. In domains such as manufacturing, automotive product configuration and mass customization, these techniques help identify combinations of components or attributes that tend to recur across different observations. Such patterns often reflect implicit dependencies, compatibility constraints or behavioural tendencies that are not immediately visible when analysing variables in isolation, a point consistently highlighted in surveys reviewing the practical relevance of association rule mining across industrial domains [17].

The core idea is to extract relationships of the form "when element X is present, element Y tends to appear as well", providing a compact representation of the structural or behavioural regularities embedded in the data. These rules are particularly valuable when the system under study involves many optional components or features that can be combined in numerous ways, making direct

modelling of all interactions impractical.

Association rules can be incorporated into predictive models by transforming them into additional features that encode the presence, strength or relevance of the discovered relationships. This strategy has been shown to enhance downstream learning tasks by injecting structured knowledge about interactions that are not explicitly represented in the raw data [18].

This enriched representation allows machine learning algorithms to capture interactions that are not explicitly encoded in the raw dataset, improving their ability to model complex configuration patterns or preference structures.

In the context of this thesis, association rule-based features are used to represent latent dependencies among choices and configurations, enhancing the modelling of interactions between variables and contributing to more robust forecasting performance.

3.6 Technology Stack

The implementation of the forecasting pipeline was carried out in Python, using a scientific computing environment designed to support reproducibility and efficient experimentation. The workflow relies on widely adopted libraries for data manipulation, modelling and evaluation, ensuring consistency with established practices in machine learning research.

The main components of the technology stack include:

- Data processing and numerical computing: `pandas`, `numpy` and parallelization utilities for efficient handling of tabular data and large-scale preprocessing tasks.
- Visualization and exploratory analysis: `matplotlib`, `seaborn` and `plotly` for descriptive analytics, diagnostic plots and interactive inspection of model behaviour.
- Machine learning and model development: `scikit-learn` for classical algorithms and evaluation utilities, `xgboost` for gradient-boosted decision trees and `tensorflow.keras` for neural network architectures.
- Time-series exploration: `statsmodels` for decomposition and preliminary analysis of temporal patterns.
- Feature interpretation: permutation-based importance methods to assess the contribution of individual predictors.

- Synthetic data generation: tools from the `sdv` ecosystem to construct statistically coherent artificial samples.
- Pattern mining: `mlxtend` for extracting association rules used to enrich the feature space.

This combination of tools provides a flexible and scalable environment for developing, evaluating and interpreting the forecasting models presented in this work.

Chapter 4

Development

Forecasting the demand for optional components plays a crucial role in production planning and supply chain management. Optional features represent a significant portion of the value associated with each motorcycle configuration, yet their demand is highly variable and strongly influenced by customer preferences, dealer behaviour and model-specific characteristics.

As described in Chapter 2, the shift to configurability at purchase time makes precise take rate forecasting essential for production continuity and supply chain efficiency.

In this context, the final objective is to create a machine learning forecasting framework to automate and support a long horizon prediction of the take rate for each optional component, that is the proportion of motorcycles expected to include a given optional, rather than predicting absolute volumes directly.

The objective of this project is therefore to design a machine learning forecasting framework capable of anticipating the monthly demand for each optional component by leveraging Ducati's historical order data.

Developing such a system requires addressing several practical and methodological challenges. First, the dataset is characterised by heterogeneous information coming from different internal sources, with varying levels of completeness and consistency. Second, many optional components exhibit sparse or irregular adoption patterns, which complicates both model training and evaluation.

Crucially, the forecasting solution must ultimately support a twelve months horizon, as such a long planning window is typically required in manufacturing environments that may have configurable products, long supplier lead times and complex production planning processes. However, the available historical depth was insufficient to directly train long horizon models. For this reason, the development process began with a short term one month forecasting setup,

used to compare multiple candidate models and establish a robust baseline.

Only after this initial phase, and through the introduction of synthetic temporal augmentation, it became possible to extend the forecasting horizon to twelve months in a reliable and operationally meaningful way.

This chapter describes the methodological development undertaken to address these challenges and construct a forecasting pipeline tailored to operational context typical of motorcycle and automotive manufacturers.

The work involved the integration of multiple data sources, the formulation of a predictive structure that models customer choices at the individual order level and the design of a modeling strategy capable of supporting the final objective of producing reliable twelve months forecasts.

Since the available historical depth did not allow for direct long horizon training, the development proceeded in a structured sequence: first establishing a one month baseline to compare candidate models, then introducing synthetic temporal augmentation to compensate for the limited data and finally extending the forecasting horizon to twelve months once the necessary temporal depth had been reconstructed.

The chapter is organized to reflect the sequential development of the forecasting framework. First, Section 4.1 describes the data collection process and outlines the main limitations encountered when working with real-world industrial data. Then, Section 4.2 details the preparation and preprocessing steps, including exploratory analyses aimed at understanding the structure of the dataset and identifying relevant patterns. Section 4.3 introduces the problem formulation, explaining how optional component demand is modeled at the order level and how the forecasting task is framed within a machine learning setting. Next, Section 4.4 presents the one-month forecasting approach, which serves as the initial baseline used to compare candidate models and establish a reliable predictive structure. After that, Section 4.5 introduces the strategy of forecasting enhanced with synthetic data. Building on this, Section 4.6 then extends the forecasting horizon to twelve months, leveraging the augmented temporal structure to support a long horizon production planning. Finally, Section 4.7 explores the integration of features based on association rule, which enrich the model with information on co-occurrence patterns among optional components and further strengthen the forecasting pipeline.

Overall, this chapter lays the foundation for the experimental analyses presented in Chapter 5, where the performance of each methodological component is evaluated and compared.

4.1 Data collection

The data for this project was primarily obtained from internal Ducati databases and provided in Excel format through the company's data management infrastructure. These datasets formed the foundation for the analysis and forecasting activities. The main sources included:

- Orders dataset: contained detailed information about past motorcycle orders. Each record included attributes regarding information about the bike model, order specific information, dealer and customer details, dates, features, and, most importantly, the optional features selected in each order.

  This dataset represents the core foundation of the forecasting framework, as all predictions are ultimately derived from order-level information. Additional datasets used throughout the project were merged onto this structure to enrich it with further attributes, but never replaced or removed any of the original order-level information.

- Monthly sales volumes dataset: provided monthly outflow, useful for understanding overall demand dynamics.

- Planning volumes (forecast) dataset: contained information about monthly forecasts of expected outflow volumes. This dataset was not merged during preprocessing but was later used to compute the optional take rate. This was done by dividing the number of optional selections per supermodel and month by the corresponding forecasted bike volume, as described in Equation 3.1.

In addition to these sources, there were other potentially valuable datasets that could not be included due to unavailability, access limitations or the fact that they were not available in a structured form. These included external data such as special events (fairs, new model launches, regulatory changes, races and global market trends) and internal data like marketing campaigns and promotions (dates, duration, type and performance results).

Incorporating these factors could have significantly improved forecasting accuracy by capturing external and internal influences on demand. Events and campaigns often create sudden spikes or drops in orders and their absence means the model relies solely on historical patterns and budget data, limiting its ability to anticipate market changes.

A summary of all internal and external data sources considered in this project is provided in Table 4.1.

Table 4.1: Summary of internal and external data sources and their contribution to the analysis

Data topic	**Source**	**Availability**	**Used**	**Utility**
Bike order history	Internal	Available	Yes	Core foundation for analysis. It provides bike model informations, dealer/customer context, dates features and optional features selected, enabling forecasting of optional selections by month.
Monthly sales volumes	Internal	Available	Yes	Useful as a baseline for demand planning of bikes and for aligning historical patterns with expectations.
Planning volumes (forecast)	Internal	Available	Partially	Not merged in preprocessing but essential to compute optional take rate (see Eqution 3.1).
Special events (fairs, launch events, races)	External	Not available	No	Capture demand shocks and spikes around events. It can improve forecasting by explaining deviations from historical trends.
Global market trends (consumer sentiment, seasonality)	External	Not available	No	Provide broader context for demand cycles and cross-market correlations to help anticipate changes not visible in internal data.
Marketing campaigns & promotions (dates, duration, type, performance)	Internal	Not available	No	Capture the impact of promotional actions on customer purchasing behaviour, enabling the model to better distinguish promotional effects from underlying demand trends.

Configurability was introduced across multiple models, each with its own set of admissible option combinations and technical constraints. In order to conduct a rigorous and interpretable analysis, this thesis and all subsequent experiments

focuses on a single model. Restricting the scope in this way ensures that the evaluation is grounded in a dataset that is sufficiently large, internally coherent and representative of real customer behaviour, thereby reducing the risk of drawing conclusions from sparse, unstable or highly heterogeneous observations.

This separation is not merely a modelling convenience but a methodological necessity. The configurability rules vary substantially across models, in fact optional bundles that are perfectly valid for one may be technically incompatible or commercially unavailable for another. Running a unified experiment across all models would therefore generate configurations that do not correspond to any feasible product in the actual configurator, undermining both the realism of the modelling assumptions and the validity of the resulting performance metrics. By isolating a single supermodel, the analysis remains fully aligned with operational constraints and preserves the fidelity of the forecasting and classification tasks.

The datasets were provided as Excel pivot tables and then converted into CSV files for processing in Python using Visual Studio Code. During this step, only the relevant columns were selected to reduce complexity and focus on attributes essential for forecasting and take rate computation.

4.1.1 Data limitations

Although the available datasets provide a solid foundation for the development of the forecasting framework, several limitations must be acknowledged. First, the absence of external and internal contextual information, such as market events, regulatory changes, promotional campaigns or marketing initiatives, represents a significant constraint. Without these factors, the model may struggle to capture sudden fluctuations in demand, reducing the accuracy and robustness of the resulting forecasts.

A further limitation is due to the recent introduction of configurability during the motorcycle ordering process. Since this functionality was launched only recently, the historical order data available for analysis begin precisely at the moment configurability became active for each supermodel.

As a consequence, the initial portion of the time series naturally exhibits a low volume of orders, which progressively increases as configurability becomes more widely adopted.

A longer historical record would not only provide a more stable and representative dataset but would also allow the model to capture seasonal patterns and recurring behaviours that are currently impossible to infer.

Additionally, in a real industrial scenario where the company can introduce new configurability features, optional components or even entire motorcycle models, such launches can generate pronounced peaks in the order flow. With a limited historical window, the model cannot reliably distinguish between structural spikes triggered by product introductions and genuine shifts in underlying demand. This lack of temporal depth reduces the model's ability to generalise and increases the risk of misinterpreting rollout related anomalies as meaningful trends.

Overall, these limitations highlight the importance of continuously updating the dataset as new months of configurability data become available. Expanding the historical horizon will progressively improve the stability, interpretability and predictive reliability of the forecasting models.

4.2 Data preparation and preprocessing

Since the objective of this study is to develop predictive models, the data required strict preprocessing and encoding to meet model specific input constraints. Forecasting and classification techniques typically require structured numerical inputs therefore, categorical, temporal and identifier variables had to be transformed into appropriate representations.

To ensure consistent data preprocessing, a dedicated Python pipeline was developed to perform data standardization, cleaning and transformation prior to model development. Special attention was paid to avoid misleading numerical encodings, such as treating categorical or binary variables as continuous floating-point values, which could negatively affect model interpretation and performance. The process involved several key steps.

Format standardization All date-related columns were converted to a consistent `datetime` format using `pd.to_datetime`. This step was necessary to ensure compatibility across datasets and to enable a reliable merge between the Orders dataset and the Monthly sales volumes dataset, which share the date dimension as a key attribute. In particular, the date column was parsed from the `YYYY %b` format (e.g. `2022 Jan`) into a standardized date type. Date standardization was also essential to avoid inconsistencies in temporal aggregation and to guarantee coherent time analyses. No time zone adjustments were required, as all dates were assumed to be in local time.

Handling missing values Records with missing identifier were removed to preserve data integrity, as this variable was critical for order level analysis. Columns exhibiting more than 60% missing values were excluded from the dataset, this threshold was selected as a pragmatic trade-off between data retention and data quality, in order to avoid relying on highly incomplete variables that could introduce noise or bias into the analysis. No explicit imputation strategy was applied to the remaining missing values, as their proportion was limited and imputing them could have introduced artificial patterns not supported by the underlying data.

Duplicate removal Fully duplicated rows, such as records identical across all columns, were identified and removed using `drop_duplicates()` to prevent redundancy and avoid potential distortions in volume counts and aggregation results.

Outlier detection The Interquartile Range (IQR) method was applied to selected numerical variables to identify potential outliers. Distributional patterns were further examined through boxplots. Since no values were identified as extreme or inconsistent with the underlying business context, no additional outlier handling was performed.

Dataset merging The Orders dataset was merged with the Monthly sales volumes dataset using a common month–year time index derived from a date column. This temporal alignment was introduced to ensure consistency between order level information and monthly sales volumes, and to enable coherent aggregation at the monthly level, which represents the shared temporal granularity across datasets. The merge followed a many-to-one logic, as multiple order records could be associated with the same month-level outflow volume value, and a left join strategy was adopted to preserve all order observations, even in cases where budget information was missing for specific periods. The Planning volumes (forecast) dataset was intentionally excluded from the preprocessing stage and reserved for subsequent take rate calculations, in order to avoid information leakage and to maintain a clear separation between historical explanatory variables and forecast-based inputs.

Feature engineering Several additional features were derived from the date-related variables to enrich the temporal representation of the data. These included calendar attributes such as year, month, day of the week and a binary weekend indicator, aimed at capturing seasonality, weekly demand patterns or to quantify delays.

Encoding and aggregation The optional configuration space exhibits substantial heterogeneity, as in a typical motorcycle-manufacturing context optional groups varying widely in size and composition across motorcycle models and product lines. As a result, once transformed into binary indicators, the optional related features form a high dimensional and sparse representation. This structural characteristic justifies the adoption of One-Hot Encoding and dedicated feature engineering strategies to ensure that the resulting feature space remains interpretable and suitable for predictive modelling.

Categorical variables with limited cardinality were transformed using One-Hot Encoding to preserve their categorical nature without imposing artificial ordinal relationships. High-cardinality categorical variables were instead encoded using Label Encoding in order to limit the dimensionality of the feature space. Following the encoding process, the data were aggregated at the order level by grouping on order identifiers. For binary optional indicators, the maximum value was retained to indicate the presence or absence of each optional within a given order.

4.3 Problem formulation

The primary objective of this project is to create a machine learning forecasting framework to support the prediction of the optional take rate over a long horizon, such as twelve months, in order to support production and supply chain planning.

To achieve this goal, the forecasting problem was addressed through a two step approach combining order level prediction and temporal aggregation:

1. **Order level binary classification**:

   Rather than directly forecasting monthly quantities, the task was first formulated as a binary classification problem. For each order and for each optional component, a dedicated classifier was trained to predict whether the optional is included (1) or not (0) in the order.

   This formulation allows the model to fully exploit all information available at the individual order level and to explicitly capture customer optional selection behaviour. Since each optional exhibits its own frequency, distribution and selection dynamics, a separate binary classifier was trained for every optional, ensuring that the modelling process remains tailored to the specific characteristics of each component.

   Logistic Regression was also evaluated to obtain calibrated class probabilities. This probabilistic view helps highlight how rare many optional

components are: for several optionals, the predicted probabilities remain low because they appear only rarely in the historical data, while others receive consistently higher values. High-confidence predictions may reflect genuine data availability for those optionals, but they may also signal possible overfitting.

2. **Temporal aggregation and take rate computation**:

   Once the order level predictions were obtained, the next step consisted of aggregating the predicted presences by the optional identifier and by month of order creation. Summing the predicted binary outcomes across all orders in a given month yields an estimate of the monthly quantity of each optional component.

   This aggregation step is essential because the take rate is defined at the monthly level and cannot be computed directly from individual order predictions. By comparing the predicted monthly volume of each optional with the total number of motorcycle orders in the same period, the corresponding take rate can be derived.

   This transformation from order level predictions to monthly quantities represents the final stage of the methodology and directly enables the objective of the project: forecasting the optional take rate over time.

4.3.1 Modeling strategy

A staged modeling strategy was adopted to identify the most suitable approach for the final twelve months forecasting task.

Initially, multiple machine learning algorithms were evaluated on a one month prediction horizon (Section 4.4). After selecting the best performing model, the historical dataset was augmented using synthetic data generation techniques (Section 4.5) in order to mitigate data sparsity issues. The selected methodology was then extended to the full twelve months forecasting horizon (Section 4.6). Finally, association rule mining was applied to better capture frequent optional combinations, and the forecasting process was repeated using the enriched feature space (Section 4.7).

The following algorithms were evaluated for the binary classification task:

- **Random Forest**: chosen for its robustness to overfitting, ability to handle mixed data types with limited preprocessing, and interpretability through feature importance measures.
- **XGBoost**: selected for its strong performance on structured data and its ability to capture complex non linear interactions via gradient boosting.

- **Support Vector Machine**: included due to its effectiveness in high dimensional spaces and its capability to model non linear decision boundaries through kernel functions.

- **Keras Neural Network**: evaluated to explore the potential of deep learning models in capturing complex and non linear patterns in the data. However, due to the limited size of the available historical data, expectations regarding its performance were relatively low, as neural networks typically require large amounts of data and careful hyperparameter tuning to fully express their modeling capacity.

- **Logistic Regression**: included to obtain calibrated class probabilities and to provide an interpretable reference for assessing class separability. Its probabilistic outputs offer insight into the inherent rarity of many optional components and help evaluate whether high confidence predictions reflect genuine patterns or potential overfitting.

To optimize model performance, a systematic hyperparameter tuning procedure was applied. Specifically, GridSearchCV was used to explore combinations of hyperparameters of the different models (e.g. number of estimators, maximum depth, minimum samples per split). The search was combined with 5-fold cross validation to evaluate each hyperparameter configuration on different training-validation splits and select the combination that maximized average accuracy.

Statistical time series approaches such as ARIMA were not considered, as the number of monthly observations available after aggregation is too limited to support reliable temporal modelling. - Given the limited historical depth available, the series do not exhibit stable seasonal patterns, long term trends or sufficient temporal depth to estimate autoregressive and moving average components in a meaningful way. Moreover, the presence of structural discontinuities caused by configurability rollouts further reduces the suitability of classical time series models, which typically assume stationarity or at least gradual temporal evolution.

Nevertheless, as additional months of configurability data become available and the historical horizon expands, time series methods could become a viable alternative. A longer time series would allow the identification of recurring seasonal behaviours and more robust temporal dependencies, enabling the application of traditional forecasting models in future work.

4.3.2 Model evaluation

Model performance was evaluated at two complementary levels, reflecting both order level prediction accuracy and aggregated forecasting performance:

1. **Order level evaluation (binary classification)**: models performances were assessed using standard classification metrics:

   - **Accuracy**, measuring the proportion of correct predictions over the total number of orders.
   - **Precision**, indicating the proportion of correctly predicted present optionals among all optionals predicted as present.
   - **Recall**, measuring the ability of the model to correctly identify optionals that are actually present in the orders.
   - **F1 Score**, defined as the harmonic mean of precision and recall, providing a balanced measure of model performance, particularly in the presence of class imbalance.

   To enhance model interpretability and error analysis, the following additional outputs were considered:

   - **Permutation Importance**, used to assess the contribution of each input feature by measuring the decrease in model performance when the feature values are randomly permuted.
   - **Confusion Matrix**, providing a detailed breakdown of true positives, true negatives, false positives, and false negatives.

2. **Aggregated monthly evaluation**: after aggregating order level predictions, predicted monthly quantities were compared against actual observations using regression and cumulative error metrics:

   - **MAE** (Mean Absolute Error), measuring the average absolute difference between predicted and actual quantities.
   - **MAPE** (Mean Absolute Percentage Error), indicating the average prediction error in percentage terms relative to actual values.
   - **SMAPE** (Symmetric Mean Absolute Percentage Error), normalizing the error with respect to the sum of predicted and actual values. This metric is more stable across optionals of different volumes.
   - **MSE** (Mean Squared Error), calculating the mean of squared errors, which penalizes larger deviations more heavily.

- **RMSE** (Root Mean Squared Error), the square root of MSE, providing an error measure on the same scale as the original data.
- **R^2** (Coefficient of Determination), indicating the proportion of variance in the actual quantities explained by the model.
- **CPE and CVE**, although not applicable in this study because component-level price information was not usable, Cost per Error (CPE) and Cost per Volume Error (CVE) would be highly relevant in an operational context. These metrics quantify the financial impact of over, or under, forecasting specific optionals, and are commonly used in supply chain optimisation to prioritise accuracy where misalignment is most costly.

This two-level evaluation framework provides a comprehensive assessment of model performance, capturing both the accuracy of predictions at the individual order level and the reliability of aggregated monthly forecasts. By doing so, it ensures that the proposed approach is not only robust in modeling customer optional selections but also suitable for informing operational planning and supply chain decisions.

In the the broader context of automotive manufacturing sector, forecasting errors have asymmetric consequences because shortages and surpluses do not carry the same operational impact. Overestimating the demand for an optional (oversampling) results in excess inventory, which increases storage and capital holding costs but does not disrupt production. In contrast, underestimating demand (undersampling) can lead to stockouts of critical components.

Table 4.2: Model Evaluation Framework

Evaluation Level	**Purpose**	**Metrics / Outputs**
Classification	Determine, for each order and each optional, whether the optional is present or absent, allowing the model to capture individual selection behaviour	Accuracy, Precision, Recall, F1 Score, Permutation Importance, Confusion Matrix
Monthly aggregation	Aggregate predicted presences by optional and month to estimate monthly quantities, enabling the computation of take rates	MAE, MAPE, SMAPE, MSE, RMSE, R^2

4.4 One month forecasting approach (baseline)

The one month forecasting approach represents the baseline methodology adopted in this project and serves as a reference point for all subsequent extensions. The primary goal of this phase is to evaluate whether optional component demand can be reliably predicted by modeling customer choices at the individual order level, using a short one month forecasting horizon. In addition, this phase aims to compare multiple candidate models and identify the most suitable one to be employed as the core forecasting engine in more complex methodological extensions, as the twelve months forecasting.

A one-step-ahead temporal evaluation setup was adopted:

- Training set: all customer orders observed up to the penultimate month;
- Test set: all customer orders belonging to the most recent available month.

This configuration reflects a real world operational scenario, in which forecasts for the upcoming month are generated using all information available up to the present time.

Optional component demand was modeled at the individual order level by framing the problem as a binary classification task applied to each order–optional pair. For every customer order, the model predicts whether a specific optional is included (1) or not (0), treating optional selection as the outcome of individual purchasing decisions. After the binary classification step, predicted optional presences are aggregated over time in order to transform order level outputs into monthly quantities. Specifically, predicted binary outcomes are summed across all orders belonging to the same month for each optional component, yielding an estimate of monthly demand.

An alternative approach would have been to directly applying a regression model to monthly aggregated quantities, but it would require collapsing all orders within the same period into a single observation, resulting in a substantial loss of information. Several explanatory variables, including customer characteristics, dealer attributes and motorcycle configurations, vary at the order level and cannot be adequately preserved through temporal aggregation, limiting the model's ability to learn individual purchase behaviour.

The adopted two stage approach mitigates this issue by modeling optional selection decisions at the order level and performing temporal aggregation only as a subsequent step. This structure allows the exploitation of the full

historical order features, significantly increasing the number of training samples compared to a purely time series formulation. Moreover, it establishes a flexible framework that can be extended to longer forecasting horizons and enhanced feature representations, as explored in subsequent sections.

4.4.1 Multi model comparison and selection

To identify the most suitable classification model for the baseline one month forecasting task, several machine learning algorithms were evaluated under a consistent experimental setup. All models were trained using the same preprocessed feature set and followed an identical pipeline in order to ensure a fair comparison.

The comparison included Random Forest, XGBoost, Support Vector Machine, Keras-based Neural Network and Logistic Regression. For each algorithm, hyperparameters were optimized using grid search combined with 5 fold cross validation applied to the training data. This procedure enabled the selection of hyperparameter configurations that provide robust performance across multiple validation splits while mitigating the risk of overfitting to a single partition.

A detailed overview of the models considered and the associated hyperparameter grids is reported in Table 4.3.

Model evaluation was conducted at both the classification and aggregation levels. At the order level, standard classification metrics such as Accuracy, Precision, Recall and F1 Score analysis were considered to assess the quality of binary predictions. In parallel, the aggregated monthly quantities derived from these predictions were compared against observed values using regression style metrics, including MAE, MAPE, SMAPE, MSE, RMSE and R^2. This two level evaluation framework allows consistency between individual order predictions and the final business relevant quantities to be assessed.

Several challenges emerged during this phase:

- A significant issue was the inherent sparsity of certain optional components, which are selected only in a limited number of orders. This class imbalance increases model variance and may lead to unstable predictions for rare options.

- Additionally, models exhibited tendencies toward overfitting, particularly when model complexity increased relative to the available training data.

Table 4.3: Machine learning models evaluated for the one month forecasting baseline and corresponding hyperparameter search spaces.

Model	**Motivation for inclusion**	**Hyperparameters explored**
Random Forest	Robust to overfitting, capable of handling heterogeneous features with limited preprocessing, and interpretable through feature importance.	n_estimators $\in \{50, 100, 150, 200\}$ max_depth $\in \{5, 10, 15, 20\}$ min_samples_split $\in \{2, 5, 10\}$
XGBoost	Strong performance on structured data and ability to capture non-linear interactions through gradient boosting.	n_estimators $\in \{50, 100, 150\}$ max_depth $\in \{3, 5, 7\}$ learning_rate $\in \{0.01, 0.1, 0.2\}$ subsample $\in \{0.8, 1.0\}$
SVM	Effective in high-dimensional spaces and capable of modeling non-linear decision boundaries via kernel functions.	C $\in \{0.1, 1, 10, 100\}$ gamma $\in \{0.1, 1, scale, 10, 100\}$ kernel = rbf max_iter = 1, cache_size = 200
Keras Neural Network	Explored to assess the potential of deep learning models in capturing complex patterns, despite limited data availability.	Neurons $\in \{32, 64\}$ Dropout rate $\in \{0.2, 0.3\}$ Batch size $\in \{32, 64\}$ Epochs $\in \{20, 50\}$
Logistic Regression	Used as a simple, interpretable baseline model that provides probabilistic outputs. These calibrated probabilities make it useful for understanding how likely each optional is to be selected.	C $\in \{0.01, 0.1, 1, 10\}$ penalty $\in \{l1, l2\}$ solver = liblinear class_weight = balanced

- Computational cost also represented a non negligible constraint, especially for Support Vector Machines and the Neural Network model. The combination of high dimensional feature space and extensive hyperparameter grids resulted in long training times, limiting the practicality of these approaches in an operational setting. In contrast, tree-based models such as Random Forest and XGBoost demonstrated a more favorable balance between predictive power and computational efficiency.

Based on the overall evaluation across classification robustness, aggregation level performance, interpretability and computational feasibility, Random Forest was selected as the reference model for the one month forecasting baseline. This model serves as the foundation for the subsequent methodological extensions presented in the following sections, including synthetic data augmentation (Section 4.5), extended forecasting horizons (Section 4.6) and feature enrichment through association rule mining (Section 4.7).

4.5 Synthetic data-enhanced forecasting

One of the main challenges encountered in the development of the take rate forecasting framework concerned the limited availability of historical data. At the time of the analysis, the order database covered less than one year of observations, a time span that is inherently insufficient for training and validating forecasting models over extended horizons, especially in the presence of strong seasonality and heterogeneous customer configurations.

This limitation is particularly critical because the business task requires evaluating model performance over a twelve months horizon. With only twelve months of real history, a model trained exclusively on observed data would lack the temporal depth needed to generalise to an entire future year.

As additional years of order data are collected and the historical archive becomes sufficiently long and stable, the synthetic augmentation step may cease to be necessary for supporting the forecasting models.

To mitigate this limitation, a synthetic data augmentation strategy was adopted. The synthetic enhancement is applied only after the identification of the best performing baseline model. The synthetic enhancement is applied only after the identification of the best performing baseline model, so that model selection and performance evaluation remain grounded on real data only.

Synthetic data generation was performed using a Gaussian Copula synthesizer, a method that models the joint distribution of the variables by separating marginal distributions from their dependency structure, which is captured

through a Gaussian copula [16]. This approach is well suited to tabular industrial data characterised by a mix of numerical and categorical variables, complex non-linear dependencies and a high dimensional feature space.

Compared to generative deep learning approaches, Gaussian Copula synthesizers offer improved stability and a lower risk of overfitting when operating on limited datasets, while still preserving realistic correlations between features. This property is particularly relevant in the present setting, where the real dataset already shows signs of overfitting due to its restricted size.

By explicitly modelling both marginal distributions and inter-variable dependencies, the Gaussian Copula is able to generate novel yet statistically coherent combinations of optional configurations. This provides a controlled and statistically grounded way to expand the training set, avoiding the instability that more flexible generative models may introduce under data scarcity.

A key requirement of the synthetic augmentation process is the preservation of temporal causality. To avoid any form of information leakage into the future, all synthetic observations are temporally shifted to precede the original dataset. Specifically, date related attributes associated with each synthetic order are consistently translated backward by one calendar year.

This operation ensures that the augmented dataset represents an extended historical period, rather than an artificial continuation beyond the observed time frame.

Once generated and temporally aligned, synthetic observations are merged with the original dataset to form an extended historical series of approximately two years. This enlarged history is first used to train the one month forecasting model, ensuring that short horizon predictions benefit from the additional context. Only after this step is the twelve months evaluation performed, using the same model architecture but relying on the richer historical window provided by the synthetic enhancement.

The introduction of synthetic data does not modify the structure of the forecasting model nor the learning objective, it solely extends the amount of past information available during training.

The effectiveness of this strategy relies on the assumption that the dependency structure learned from the available data is representative of the true underlying process.

While the Gaussian Copula synthesizer preserves global statistical properties, it may smooth rare events or underrepresent abrupt structural changes. Moreover, the limited availability of real historical data remains a fundamental

constraint. In fact, with only the first year of configurability available, the synthetic augmentation inevitably reflects patterns that may not yet be fully mature and that correspond to a specific set of optionals whose composition is likely to evolve over time.

For these reasons, synthetic data are treated strictly as a supporting mechanism rather than a source of ground truth, and all quantitative performance assessments are conducted exclusively on real observations.

4.6 Twelve month forecasting approach

After enriching the historical dataset through synthetic augmentation, the next step of the pipeline consists of evaluating the model's ability to generalize to an entire future year.

The twelve months horizon is assessed by reserving the most recent year of observations as an out-of-sample evaluation window. The model is trained on the full historical period preceding this window and then applied month by month to the held out year, providing a realistic indication of its predictive behaviour across an entire production cycle.

This configuration reflects the operational requirements typically found in motorcycle manufacturing and the structure of standard production planning processes. Evaluating model performance over a full year therefore provides a robust indication of the classifier's stability, its ability to capture seasonal and structural patterns, and its practical usefulness in a real deployment environment.

The procedure is straightforward: the dataset is ordered chronologically, the last twelve months are isolated as the test period and the model is trained on all preceding months. For each optional, the classifier is then applied to every month of the held out year, producing predictions on real, unseen observations. This month-by-month evaluation allows the analysis of temporal stability and highlights potential drifts in customer behaviour or optional availability.

The twelve months evaluation must be interpreted in light of the structural characteristics of the dataset. The configurability program is recent, optional components were introduced at different points in time and the historical series does not yet exhibit stable long term patterns. For this reason, the most recent year is treated as a held out evaluation window rather than as a forecasting horizon.

By training the model on all preceding months and assessing its behaviour across the final twelve, the analysis captures how the classifier would have performed if deployed during the latest production cycle. This framework provides a transparent and operationally meaningful assessment of predictive stability over time, without relying on assumptions about future dynamics. The quantitative results of this evaluation are presented in Section 5.3.

4.7 Association rules feature augmentation

The configurability allows customers to select multiple optional components within the same order. Understanding how these options tend to co-occur is therefore strategically relevant, as certain accessories are frequently purchased together while others appear only in specific combinations. These patterns reflect underlying customer preferences, compatibility constraints and broader marketing dynamics.

Incorporating this information into the forecasting pipeline can enrich the feature space with higher–order interactions that are not explicitly captured by the individual optional indicators.

To extract these co-occurrence patterns, an association rule mining procedure was applied to the one–hot encoded optional.

Association rules are a classical data mining technique used to identify sets of items that appear together more frequently than expected under independence. Given a binary transaction matrix, the Apriori algorithm identifies frequent itemsets whose empirical support exceeds a predefined threshold [19]. From these itemsets, rules of the form $X \rightarrow Y$ are generated, where X and Y are disjoint sets of items. Each rule is characterised by three key metrics:

- Support: the proportion of orders containing both X and Y;
- Confidence: the conditional probability that an order containing X also contains Y;
- Lift: the ratio between the observed co-occurrence of X and Y and the expected co-occurrence under independence.

Support measures prevalence, confidence measures reliability and lift quantifies the strength of the association beyond chance. These metrics are particularly suitable for the present application, where optional configurations are high-dimensional and sparse, and where meaningful interactions may involve combinations of multiple components.

The Apriori algorithm was applied to the full set of optional using a minimum support threshold of 1%, meaning that only itemsets appearing in at least 1% of all orders were retained. This constraint prevents the generation of rules based on extremely rare combinations and ensures that the resulting patterns are statistically meaningful.

This yielded a total of 108,809 association rules. To ensure interpretability and avoid an excessive expansion of the feature space, only the most informative rules were retained. Specifically, the rules were ranked by confidence and lift, and the top twenty were selected for feature augmentation. For each selected rule, a new binary feature was created, indicating whether all antecedent and consequent items of the rule were jointly present in a given order. These feature combos capture higher–order interactions that cannot be represented by individual optional indicators alone.

The newly engineered features were merged into the original dataset and used to construct an augmented version of the historical series. The twelve months evaluation described in Section 4.6 was then repeated using this enriched feature space.

Although the improvement in predictive performance was moderate, the results highlight the potential value of incorporating structured co-occurrence information into the forecasting model. In particular, the augmented features appear to help the classifier better capture complex configuration patterns that are characteristic of customer behaviour in the configurability programme.

This approach, however, presents several limitations. First, association rules are static and do not account for temporal evolution: the popularity of certain option bundles may change over time, especially in a context where new components are introduced and customer preferences evolve. Second, the selection of rules is based on global metrics and does not explicitly consider their predictive relevance for individual optionals. Finally, the binary nature of the engineered features may oversimplify more nuanced relationships between options.

Overall, the association rule augmentation represents a lightweight yet effective enhancement to the forecasting pipeline, providing a structured way to incorporate domains-pecific interactions between optional components.

Chapter 5

Experimental results

This chapter presents the experimental assessment of the forecasting framework designed to predict the monthly take rates of optional components for a selected supermodel.

The analysis examines how effectively the proposed methodology captures real demand patterns, how robustly it performs across different temporal segments and how well it aligns with operational constraints. By structuring the evaluation around a single supermodel, the chapter provides a focused and coherent view of model behaviour, enabling a clear interpretation of forecasting accuracy, stability and practical applicability within a real production planning context.

The results discussion covers the four main experimental stages previously described: baseline forecasting over a one month horizon (Section 5.1), forecasting enhanced with synthetic data over a one month horizon (Section 5.2), long term twelve months forecasting (Section5.3) and finally the extraction of association rules and their impact on model performance when used as feature augmentations (Section 5.4).

All experiments were conducted under realistic computational constraint, and particular attention is given to model robustness, generalization behaviour and the interpretability of results.

The computational cost varied significantly across machine learning models and experimental settings. For the one-month forecasting task, training times ranged from lightweight models such as SVM (20 minutes) and Logistic Regression (25 minutes), to more computationally demanding architectures such as Keras neural networks (190 minutes). Random Forest and XGBoost required approximately 30 and 40 minutes respectively, offering a favourable balance between accuracy and efficiency.

These differences emphasise the importance of selecting models that balance predictive performance with computational feasibility, especially in operational environments where frequent model updates are required.

To evaluate forecasting performance, model outputs were aggregated by optional and month. Since the underlying task is multi-label classification at the order level, the monthly forecast is obtained by summing the predicted selections for each optional. This procedure effectively transforms the classification task as a monthly quantity regression problem, enabling the use of standard forecasting metrics such as MAE, RMSE and R^2.

See Chapter 4.3.2, introduced earlier, for a detailed description of the evaluation methodology.

Error analysis revealed that the largest discrepancies occur for optional with extremely low or extremely high demand. Rare options amplify percentage errors (MAPE, SMAPE), while highly frequent options dominate the aggregated error contribution. This behaviour reflects the intrinsic long tailed structure of the dataset, where a small number of highly frequent optionals dominate the distribution, while the majority of options appear only sporadically. As a consequence, percentage error metrics tend to be inflated for rare options, whereas frequent options contribute disproportionately to the aggregated error.

5.1 Baseline forecasting results

The baseline experiment assesses the performance of several machine learning models over a one month forecasting horizon. The prediction task involves a sparse and highly multi-label dataset in which each optional is modelled independently and many optionals appear only a few times per month. This combination of low support and strong heterogeneity makes the forecasting problem particularly challenging and amplifies the differences in model behaviour under data scarcity.

This structure increases the risk of overfitting, particularly for high capacity models such as neural networks and gradient boosting, and limits the predictive feasibility for optional with insufficient historical support. Overfitting was observed in the discrepancy between training and test metrics.

Before training the forecasting models, a minimum volume threshold was applied to exclude optionals with insufficient historical support.

In the original dataset, several optional appeared extremely rarely, sometimes only one or two times in the entire month. These low frequency optionals

included, for example, particular color schemes or configurations introduced late in the production cycle.

To ensure model stability, a threshold of at least 20 occurrences per month was adopted. Optional appearing in fewer than twenty orders were excluded from model training and evaluation. This filtering step was necessary for two main reasons:

- Algorithms such as Logistic Regression, SVM and Neural Networks require a minimum number of positive samples to construct a meaningful decision boundary. When the positive class is extremely rare (e.g. 1 positive sample out of hundreds of orders), these models either fail to converge, produce degenerate solutions or collapse into predicting only the negative class.

- With only few historical observations, no model, regardless of complexity, can generalize a meaningful pattern. Including such optionals would artificially inflate error metrics (especially MAPE and SMAPE) and introduce noise into the aggregated evaluation.

Applying this threshold reduced the number of modelled optionals more or less from 41 to 27 (approximately 34% of the rarest optionals were excluded). The excluded optionals were primarily niche configurations or accessories with very limited adoption, whose forecasting would not have provided operational value due to the absence of sufficient historical signal.

This preprocessing step ensured that the baseline comparison across models was conducted only on optionals for which a meaningful binary classification task was feasible.

Despite these challenges, the baseline results provide a meaningful reference point for evaluating subsequent improvements.

Table 5.1 reports the aggregated monthly prediction metrics across all optionals. The comparison highlights substantial variability in performance across models, especially in terms of MAPE and SMAPE, which are strongly affected by low-volume optionals.

Although XGBoost and the neural network achieve slightly lower RMSE values, their performance is less stable across optionals and more sensitive to data sparsity. Logistic Regression, as expected, performs poorly due to the strong nonlinearities in the data.

Table 5.1: Comparison of aggregated monthly prediction metrics across baseline one month forecasting models.

Model	MAE	MAPE	SMAPE	MSE	RMSE	R^2
RF	10.19	20.78	67.80	353.58	18.80	0.978
XGB	10.05	22.10	49.20	348.10	18.65	0.973
SVM	10.11	17.20	49.88	341.22	18.47	0.969
Keras NN	9.32	41.55	52.90	289.77	17.02	0.979
LR	32.44	884.11	52.77	3104.55	55.73	0.732

To illustrate the overfitting issue, Table 5.2 reports the Random Forest performance on both training and test sets. The discrepancy between the two is substantial, the model fits the training data almost perfectly, while the test error remains significantly higher.

A similar pattern is observed across the other models evaluated in the baseline experiment, although the magnitude of the gap varies depending on the model architecture and the amount of data available for each optional.

This behaviour was expected given the limited number of observations per optional and the high dimensionality of the feature space. Nevertheless, the Random Forest model remains competitive and comparatively more stable than other high capacity models.

Table 5.2: Random Forest aggregated monthly prediction metrics on the training and test sets (baseline configuration).

Model	MAE	MAPE	SMAPE	MSE	RMSE	R^2
Train	0.423	2.436	7.112	2.680	1.636	0.999
Test	10.192	20.777	67.798	353.577	18.804	0.978

Although models such as XGBoost and Neural Networks achieved slightly lower error metrics in some configurations, Random Forest was selected as the primary model for the subsequent forecasting analysis. This choice is motivated by both empirical evidence and practical considerations.

- Robustness under data sparsity and evolving feature space:
  At the beginning of the project, the dataset contained only six months of orders, a volume that strongly limited the learning capacity of high complexity models. Under these conditions, Random Forest consistently

outperformed all other models, providing stable predictions even for optionals with very few observations.

As additional months of historical data were incorporated, models such as XGBoost and Neural Networks improved, but their performance remained highly dependent on the specific optional and on the richness of its historical series. In practice, these models performed well only for optionals with sufficient historical support, while their accuracy fluctuated considerably for low-volume or newly introduced codes.

Furthermore, the configurability space evolved over time: new optional were introduced, others discontinued and some underwent structural changes. These dynamics altered both the feature distribution and the target space, often destabilising models that rely on finely tuned hyperparameters or strong assumptions about data regularity. Random Forest demonstrated the highest resilience to these changes, maintaining stable performance across different phases of the project. In contrast, models such as XGBoost and Neural Networks exhibited more pronounced sensitivity to these variations, requiring additional tuning or suffering from inconsistent generalisation.

This consistency, combined with its early superiority under data scarcity, made it unnecessary to revise the modelling choice later on, even when other models achieved slightly lower error metrics in specific settings.

- Lower sensitivity to overfitting:
  While all models exhibited some degree of overfitting, Random Forest showed the most predictable and controllable behaviour. High capacity models tended to overfit more aggressively. Random Forest, instead, maintained strong generalization performance across optionals.

- Interpretability and operational transparency:
  Feature importance measures derived from Random Forest provide actionable insights into the drivers of optional take rates. This interpretability is essential in a configurability pipeline, where understanding the contribution of each feature supports both business decisions and model validation.

- Computational efficiency and scalability:
  The forecasting pipeline requires training one classifier per optional. Given that the number of available optional is large and that each model must be retrained whenever new orders become available, training time and computational cost become critical operational factors.

In this context, Random Forest exhibits several properties that make it a particularly effective choice for addressing these operational challenges:

- Fast training and retraining cycles, in fact the model can be fitted quickly even when repeated across dozens of optional, enabling frequent updates of the forecasting pipeline without incurring excessive computational overhead;
- Predictable behaviour under class imbalance;
- Robustness to heterogeneous feature encodings.

Since the forecasting process may need to be updated regularly, potentially multiple times per month as new orders are recorded or as production decisions evolve, computational efficiency is not merely a convenience but a key operational requirement.

Forecasting is used to support the planning of optional components and to adjust orders that are still modifiable within the production window. Whenever the forecast highlights a significant deviation from expected demand, the planning team may need to rerun the model to evaluate alternative scenarios or to update the order mix accordingly. This operational setting naturally leads to frequent retraining cycles, often triggered by new data arrivals or by the need to validate revised production strategies.

Random Forest provides a favourable balance between accuracy and speed, ensuring that the pipeline remains maintainable and responsive even under frequent retraining. Its fast fitting time, combined with its robustness to heterogeneous features and class imbalance, makes it particularly suitable for a forecasting system that must remain agile and continuously up-to-date.

- Consistent performance across metrics and time:
  Random Forest delivers a strong balance across MAE, RMSE, and R^2 and its performance remained competitive throughout the project, even as the dataset evolved. This consistency, combined with its robustness and interpretability, makes it the most suitable choice for the final twelve-month forecasting task.

In light of the empirical evidence and the practical constraints discussed above, Random Forest was identified as the most suitable model for the twelve-month forecasting task.

Random Forest offers the best trade-off between accuracy, robustness, interpretability and computational feasibility. Although alternative models,

such as Gradient Boosting and Neural Networks, occasionally achieved lower error metrics, their instability across optionals and sensitivity to data availability made them less suitable for long term, frequently updated forecasting in a real configurability environment.

For these reasons, Random Forest offers the best balance between accuracy, robustness and operational feasibility, making it the most appropriate choice for extended horizon demand prediction.

5.2 Synthetic data-enhanced forecasting results

The following section evaluates how the introduction of synthetic observations affects the short horizon (one month) forecasting performance and whether the extended temporal context provides measurable benefits or introduces new sources of variability.

To ensure that the synthetic augmentation step produced meaningful and statistically coherent observations, the generated records were compared with the real dataset along several dimensions. The Gaussian Copula model preserves the joint distribution of the original features, ensuring that global correlations, co-occurrence patterns and marginal distributions remain consistent with the real configurability space.

A qualitative inspection confirms that the synthetic samples reproduce the main structural properties of the real data:

- the distribution of categorical and numerical features remains aligned with the original dataset;
- the correlation matrix of the synthetic data closely matches that of the real observations;
- rare configurations are smoothed but not distorted, reducing the risk of overfitting while maintaining realistic variability.

These checks ensure that the synthetic records extend the historical depth without introducing artefacts or unrealistic patterns. For this reason, the augmented dataset can be reliably used to support longer horizon forecasting.

At this intermediate stage of the pipeline, the augmented dataset, composed of twelve months of real observations and twelve months of temporally shifted synthetic records, is used to retrain the one month forecasting using Random

Forest model to assess whether a richer temporal context could improve short horizon predictive accuracy.

Importantly, the evaluation remains strictly grounded on real data, in fact the test set still corresponds to the last observed month, ensuring that performance comparisons remain meaningful and unbiased.

Table 5.3 reports the aggregated prediction metrics obtained after retraining the model on the extended 24 month history.

Table 5.3: Aggregated monthly prediction metrics for the one month forecasting model trained on the dataset augmented with synthetic data.

Model	**MAE**	**MAPE**	**SMAPE**	**MSE**	**RMSE**	**R^2**
Train	4.407	25.011	80.736	124.664	11.165	0.985
Test	8.928	26.863	97.138	390.142	19.752	0.976

To contextualise these results, Table 5.2 summarises the performance of the same model when trained exclusively on the original shorter dataset.
A direct comparison reveals a clear shift in the error profile:

- Training error increases substantially: MAE from 0.423 to 4.407 and RMSE from 1.636 to 11.165.
  This behaviour is expected, the synthetic dataset introduces additional variability and reduces the risk of overfitting to the limited real observations. The model is exposed to a broader range of configurations and dependency patterns, which naturally leads to a less tightly fitted training performance.

- Test error shows a moderate increase: MAE from 10.192 to 8.928 and RMSE from 18.804 to 19.752.
  Despite the larger training set, the one month forecasting accuracy does not improve. The slight deterioration is consistent with the fact that synthetic data, while statistically coherent, cannot fully replicate the fine grained structure of real customer behaviour.
  The model therefore learns a smoother, more general representation of the underlying process, which may sacrifice some precision.

- The R^2 score remains high and stable: from 0.978 to 0.976.
  The results indicate that the model still captures the majority of the variance in the target variable even after the introduction of synthetic observations.

The shift in performance metrics reflects the intended effect of the synthetic augmentation step. The goal at this stage is not to optimize one month accuracy, this was already achieved in the baseline scenario, but to prepare the model for longer horizon forecasting, where the limited real history would otherwise be a severe constraint.

The introduction of synthetic data modifies the learning dynamics in three main ways:

1. Reduced overfitting:
   With a larger and more diverse training set, the model becomes less tightly fitted to the specific patterns present in the first year of configurability. This leads to higher training error and a more balanced gap between training and test performance.

2. Smoother dependency structure:
   Gaussian Copula synthesis preserves global correlations but tends to smooth rare or extreme patterns. As a result, the model learns a more regularised representation of option co-occurrence, which may slightly reduce short term accuracy but improves robustness.

3. Improved temporal context:
   Although the one month horizon does not fully benefit from the extended history, the enriched dataset provides the necessary foundation for the subsequent twelve months forecasting step (Section 5.3). The model now has access to a longer sequence of seasonal and configurational patterns, which is essential for predicting over an entire year.

Overall, the introduction of synthetic data leads to a controlled increase in prediction error for the one month horizon, but this effect is both expected and acceptable within the design of the pipeline.

The synthetic enhancement step is not intended to outperform the baseline short term model, rather, it enables the transition to long horizon forecasting by providing a richer historical context while maintaining stable and reliable performance on real test data.

The next section evaluates the impact of this augmented history on the twelve months forecasting task.

5.3 Twelve month forecasting results

The final stage of the synthetic data-enhanced pipeline consists of evaluating the model over a twelve months forecasting horizon.

This experiment is considerably more demanding than the one month scenario, as it requires the model to propagate uncertainty across an extended sequence of predictions while relying on a training history that, although enriched with synthetic observations, remains relatively limited.

In this scenario, the aim is not only to minimise prediction error, but also to verify that the model can sustain a twelve months forecasting horizon without loss of stability or reliability.

Table 5.4 reports the aggregated monthly metrics obtained by the Random Forest model when trained on the dataset enhanced with synthetic data and evaluated over the full twelve months horizon.

Table 5.4: Aggregated monthly prediction metrics for the twelve months forecasting model trained on the dataset augmented with synthetic data.

Split	**MAE**	**MAPE**	**SMAPE**	**MSE**	**RMSE**	**R^2**
Train	5.628	48.620	169.939	191.893	13.853	0.959
Test	7.187	46.261	166.253	321.963	17.943	0.963

Despite the increased complexity of the task, the model achieves stable and coherent performance across the twelve months horizon. The R^2 values remain high, 0.959 on training and 0.963 on test, indicating that the model continues to explain a substantial portion of the variance even when projecting one full year ahead.

The error metrics naturally increase compared to the one month scenario. This behaviour is expected and reflects the cumulative nature of long horizon forecasting:

- the model is trained on a dataset that, although extended, still contains only one year of real observations, limiting its ability to capture long range seasonal dynamics;
- synthetic data provide additional structure but inevitably smooth rare patterns and reduce the granularity of month to month fluctuations.

Interestingly, the gap between training and test performance remains narrow, suggesting that the model generalises well within the available historical

context and does not overfit to the synthetic portion of the dataset.

The twelve months task is inherently more demanding for several reasons:

- Propagation of uncertainty:

  The twelve months horizon remains inherently more uncertain because the model must estimate outcomes that lie progressively further from the observed data.

  As the temporal distance from the last real observation increases, the predictive signal becomes weaker and the model relies more heavily on patterns learned from the synthetic data-enhanced history.

  This naturally results in higher MAE, MAPE and SMAPE values compared to the short term scenario.

- Limited real historical depth:

  Even after augmentation, the model effectively learns long-term patterns from only one year of real data, restricting its ability to capture seasonal cycles or structural changes (see Section 4.5).

- Synthetic data smoothness:

  The Gaussian Copula synthesizer tends to underrepresent abrupt shifts or rare option combinations, training the model on a more regularised version of the historical process (see Section 4.5).

Overall, the model demonstrates robust long term forecasting capability, maintaining consistent performance across the twelve months horizon and achieving high explanatory power despite the limited real history available. The observed error levels are aligned with expectations for this forecasting setup and confirm that the synthetic augmentation step successfully provides the temporal depth required for year predictions, without introducing instability or degrading generalisation.

Beyond the quantitative results, this experiment highlights the potential of the proposed framework. Even when trained on a constrained historical dataset, partly reconstructed through synthetic observations, the model is able to sustain a full year forecasting horizon with stable behaviour.

This suggests that, in a future scenario where a much longer historical data will become available, the same architecture could leverage richer seasonal patterns, more complete and stable optional configuration, and a more diverse historical context to achieve substantially higher accuracy and stronger long

range reliability.

In this sense, the present evaluation not only demonstrates the feasibility of long horizon take rate forecasting under current data limitations, but also provides a solid foundation for future iterations, where an expanded real dataset is expected to unlock the full predictive potential of the methodology.

5.4 Association rules mining and feature augmentation

The final step of the analysis investigates whether structured co-occurrence patterns between optional components can enhance the twelve months forecasting model.

While the synthetic augmentation step expanded the temporal depth of the dataset, it did not explicitly encode the relational structure of customer choices. Association rules provide a complementary perspective, they capture stable configuration patterns that emerge from customer behaviour and can be incorporated as higher-order interaction features. This section evaluates the empirical impact of these rule-based features on long horizon forecasting performance.

Since the methodological aspects of rule extraction and feature construction were already discussed in Section 4.7, this section focuses exclusively on the empirical results and on how the additional feature combos influence predictive performance.

As expected, several rules involve highly frequent components, which naturally appear in a large proportion of orders. This behaviour is consistent with the distribution of item frequencies observed in the dataset, where a small number of optional components dominate the configuration space.

Figure 5.1 illustrates the distribution of all generated rules in the support–confidence space. Most rules exhibit low support, reflecting the sparsity of optional combinations, while a smaller subset achieves both high confidence and high lift.

These rules correspond to strong, non-random co-occurrence patterns and represent meaningful configuration structures that can enrich the forecasting model.

Figure 5.2 presents a confidence heatmap for the most frequent antecedent–consequent pairs. The block structure visible in the matrix highlights clusters of components that tend to co-occur, forming coherent configuration patterns. These patterns often correspond to accessory bundles or stylistic packages that customers select consistently, reinforcing the relevance of incorporating rule-based interactions into the forecasting pipeline.

The heatmap is displayed without item labels, as its purpose is not to analyse individual components but to highlight the presence of block structures and co-occurrence clusters that emerge from the rule set. The visualisation refers to the top 20 association rules, ordered by decreasing confidence

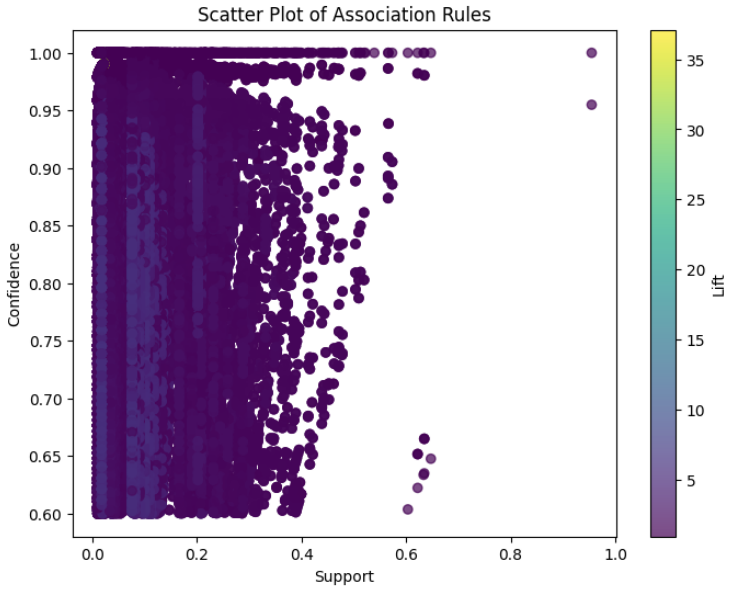

Figure 5.1: Scatter plot of association rules showing the relationship between support and confidence. Colour intensity represents lift.

From the full set of extracted rules (108809), the top twenty were selected based on their confidence values and used to construct the augmented feature set. Each selected rule was encoded as a binary feature combo, indicating whether all items in the antecedent and consequent were jointly present in a given order.

Table 5.5 compares the aggregated monthly metrics before and after the augmentation. The introduction of rule-based features produces a mixed but informative effect.

On the one hand, the error metrics increase, particularly for MAPE and

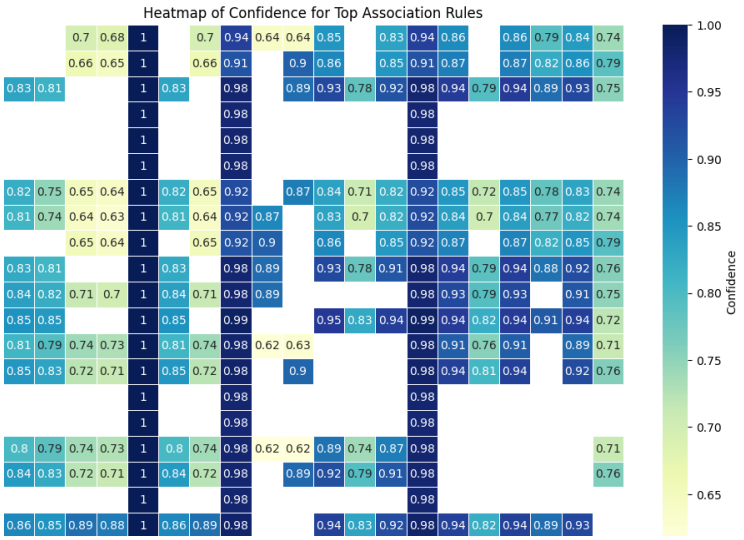

Figure 5.2: Heatmap of association rule confidence values, displayed without item labels to emphasise the underlying block structure and co-occurrence clusters.

SMAPE. This behaviour is expected as the new features capture higher-order interactions that are not always uniformly predictive across all optionals, and their binary nature may introduce additional sparsity.

On the other hand, the augmented model exhibits improved performance for several specific components and displays a more structured understanding of configuration patterns. The results suggest that association rules provide complementary information that the model can exploit, especially in cases where customer choices follow consistent co-occurrence logic.

Table 5.5: Comparison of aggregated monthly prediction metrics before and after the association rules-based feature augmentation.

Split	**MAE**	**MAPE**	**SMAPE**	**MSE**	**RMSE**	**R^2**
Base evaluation						
Train	4.407	25.011	80.736	124.664	11.165	0.985
Test	8.928	26.863	97.138	390.142	19.752	0.976
Association rules augmentation						
Train	5.145	45.974	161.501	175.718	13.256	0.962
Test	7.201	37.234	131.179	442.600	21.038	0.950

Although the overall improvement is moderate, the experiment highlights the potential of association rule-based feature engineering. Even within the constraints of a limited historical archive, the augmented model is able to leverage structured co-occurrence patterns to refine its representation of customer behaviour. As the configurability programme evolves and a richer multi-year dataset becomes available, these interactions are expected to become more informative and more stable, enabling the forecasting model to capture increasingly complex relationships between optional components.

In this sense, the present evaluation demonstrates the feasibility of integrating domain-specific interaction patterns into the forecasting pipeline and provides a foundation for future enhancements based on richer historical data and more expressive feature construction techniques.

Chapter 6

Conclusion and future work

This thesis has introduced a machine learning forecasting framework aimed at predicting the monthly take rate of optional components in configurable motorcycle models, addressing the operational complexities that characterise manufacturing contexts that have shifted toward customer-driven configurability.

The work demonstrates that, even under severe data limitations, it is possible to construct a coherent modelling pipeline that integrates order level classification, synthetic data generation and behavioural features derived from association rules. At the same time, the results confirm the expectations set at the beginning of the project, so that forecasting accuracy remains modest and signs of overfitting emerge across several models, reflecting the intrinsic constraints of the available dataset.

The experimental analysis shows that tree-based models, particularly Random Forest, provide the most stable performance across both the baseline and the extended forecasting horizons. Their ability to capture nonlinear relationships and interactions among features proves valuable in a context characterised by heterogeneous optional components and strong variability in customer behaviour. However, the limited historical depth and the sparsity of many optional components inevitably restrict the model's capacity to generalise. The gap between training and test performance, especially in the one month baseline scenario, highlights the presence of overfitting, an outcome that aligns with the structural characteristics of the data and the short observation window.

The introduction of synthetic data through a Gaussian Copula model helps mitigate this issue by extending the effective training horizon, enabling the model to sustain a long horizon prediction window with greater stability. Yet, synthetic augmentation cannot fully compensate for the absence of real temporal patterns and its benefits, while tangible, remain bounded.

The integration of association rule-based features enriches the representation of co-occurrence patterns among optional components and provides additional behavioural signals to the forecasting models. Although the overall impact on aggregated metrics is moderate, these features contribute to a more nuanced understanding of customer configuration behaviour.

These results should be interpreted considering the structural limitations of the dataset. At the time of the experiment, the historical series covers a short historical time span of real configurability data, preventing the identification of stable temporal dynamics and limiting the applicability of classical time series models. The absence of exogenous variables in a structured form, such as marketing initiatives, supply constraints or macroeconomic indicators, further restricts the model's ability to capture external drivers of demand. Moreover, the sparsity of certain optional components and the imbalance between classes make it difficult to learn reliable patterns, particularly for low frequency configurations. The static nature of the association rules also prevents the model from capturing the temporal evolution of customer preferences, an aspect that is likely to become increasingly relevant as configurability expands and personalization becomes a defining feature across many markets.

Despite these constraints, the work provides a solid foundation for future developments. As additional months of real configurability data become available, the forecasting framework can evolve toward models capable of capturing seasonality, trends and long term behavioural dynamics. The inclusion of exogenous variables would allow the system to reflect broader market conditions and operational constraints, potentially improving predictive accuracy. Finally, dynamic association rules or sequential pattern mining techniques could offer a more flexible representation of evolving customer preferences.

In conclusion, this thesis should be viewed as an initial step toward a more comprehensive forecasting system for optional components in a highly configurable manufacturing environment.

While the predictive performance is still limited and affected by overfitting, the methodological framework developed here demonstrates that meaningful insights and operationally useful forecasts can be obtained even under severe data constraints.

The approaches introduced in this work can serve as a foundation for more advanced, accurate and data-driven machine learning forecasting solutions.

Credits

Grazie a tutta la mia famiglia, agli amici e a tutte le persone che mi vogliono bene e che hanno condiviso con me questi anni.

Bibliography

[1] Nehalben Ranabhatt, Sérgio Barreto, Marco Pimpão, and Pedro Prates. Demand forecasting in the automotive industry: A systematic literature review. *Forecasting*, 7(4), 2025.

[2] Oumaima Sarhir, Zoubida Benmamoun, and Mouad Ben Mamoun. Prediction analysis for demand forcasting in automotive industry. In *2024 10th International Conference on Optimization and Applications (ICOA)*, pages 1–6. IEEE, 2024.

[3] Ikhlef Jebbor, Hanaa Hachimi, and Zoubida Benmamoun. Artificial intelligence in predicting automotive supply chain disruptions: A literature review. In Noreddine Gherabi, Janusz Kacprzyk, and Sara Arezki, editors, *Advances in Intelligent Systems and Digital Applications*, pages 11–21, Cham, 2025. Springer Nature Switzerland.

[4] Deloitte. 2024 global automotive consumer study, 2024. Accessed 2025.

[5] Iulia Novacescu. Customer insights in automotive. *Knowledge Horizons. Economics*, 11(4):20–30, 2019.

[6] Viktor Ganchev. Automotive business marketing with the use of artificial intelligence technologies for service personalization. *Актуальні питання економічних наук*, (15), 2025.

[7] Frank T. Piller and Mitchell M. Tseng. 30 new directions for mass customization. In *The Customer Centric Enterprise*, pages 519–540. Springer, 2003.

[8] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2nd edition, 2018.

[9] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5th edition, 2015.

[10] Sai Mani Krishna Sistla, Gowrisankar Krishnamoorthy, Jawaharbabu Jeyaraman, and Bhargav Kumar Konidena. Machine learning for demand forecasting in manufacturing. *Int J Multidiscip Res (IJFMR)*, 6:1–11, 2024.

[11] Real Carbonneau, Kevin Laframboise, and Rustam Vahidov. Application of machine learning techniques for supply chain demand forecasting. *European journal of operational research*, 184(3):1140–1154, 2008.

[12] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[15] Sanket Kamthe, Samuel Assefa, and Marc Deisenroth. Copula flows for synthetic data generation. *arXiv preprint arXiv:2101.00598*, 2021.

[16] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE, 2016.

[17] Qiankun Zhao and Sourav S Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 135(2003116):1–20, 2003.

[18] Feri Sulianta. Advancements and applications in association rule mining: A review of key algorithms and future directions. *(Review paper, widely cited; available on ResearchGate)*, 2023.

[19] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, pages 487–499, 1994.