Forecasting purchasing lead time by using machine learning techniques

Master's thesis in
Industrial Engineering

MOUSSAOUI Younes
HIBA Younès,

Advisor:
Dr. Ghomari

Academic year:
2022-2023

Abstract: Supply chain management (SCM) is critical for large companies seeking to optimize their efficiency, reduce costs, improve their products, and stay competitive in an increasingly complex market. Lead times can significantly impact a company's profitability. Lead time management is a complex challenge for inventory management and logistics, as shorter lead times can reduce safety stock levels, but poor management can result in additional costs and operational inefficiencies. Therefore, it is essential to manage lead times to improve a company's profitability effectively. Among the different methods for predicting lead times, machine learning methods can be accurately used to predict these times and thus reduce the costs associated with poor management. This research aims to investigate and apply machine learning techniques to predict purchasing lead times effectively.

Key-words: Lead time, Machine learning, Algorithms, supply chain, costs, profitability, inventory management , predicting .

Contents

1 Introduction 2
2 A Review of Previous Works on Purchasing Lead Time Forecasting in the Industry 2
3 Case study 8
4 Conclusions 12
5 Appendix 14

Pr G. ... of M. F. E. ... 2022/2023

1. Introduction

Supply chain management (SCM) is crucial to many large companies. The main objective is to optimize each chain step to improve overall efficiency, reduce costs, manufacture better products, and gain a competitive advantage in the market [1].

The current market is characterized by increasing demand for customized and specialized products, which has led to an evolution of the market, making it more competitive and complex. Therefore, integrating and aligning procurement with the company's strategic planning can bring competitive advantages and benefits, especially in a market with shorter lead times [2].

Among these lead times, lead times for purchased items or materials, which constitute an essential part of the supply chain and must be refined to produce the final product, represent the time between establishing a purchase order and the item's arrival at the buyer's location. In some cases, this only includes the product's delivery and sometimes manufacturing [1].These lead times are often considered an uncontrollable constant or stochastic variable since they are beyond the control of the purchasing company, which must buy them from external suppliers [3].).As a result, products are often delivered later than expected, which can have negative consequences on the supply chain [1],including significant financial losses for the company, especially in the event of order cancellation by the customer, depreciation, or obsolescence of components [4].).In addition, they can influence consumer choice and directly impact the company's sales [5].

Purchasing Lead time also plays a crucial role in inventory management in a production environment, as shorter lead times can lead to lower safety stock levels, thereby limiting potential losses for companies [3], especially in vital industries such as pharmaceuticals [1].Optimal inventory management is one of supply chain management's most complex and practical challenges. This challenge involves the periodic replenishment of perishable items with different delivery times, unit costs, and volumes in a warehouse.

Developing an optimal replenishment and lead time policy is a critical decision that directly affects the total inventory cost, including costs related to excessive or insufficient storage levels and transportation costs [6]. Indeed, poor management of lead times can increase costs and reduce operational efficiency through stock-outs, increased product counts, delays in receiving components [1], and the bullwhip effect, characterized by significant fluctuations in orders upstream of the supply chain [7].Therefore, it is essential to effectively manage Order purchasing lead times to improve the company's profitability [7].

Accurate management and forecasting of purchasing lead times can result in significant cost reductions. For example, time consolidation allows products from multiple suppliers to be grouped for delivery to a single customer. This reduces the number of vehicles needed for transportation and decreases the size of the workforce required. This method is particularly effective when the lead time is less than the average, which is the time between the customer placing the Order and the Order being received in the warehouse. By reducing transportation and labor costs, the use of time consolidation can help improve overall company profitability [8].

The main focus of this study is to explore and use machine learning methods as a suitable solution for generating accurate forecasts of purchasing lead times for large enterprises. Although machine learning has been used in various aspects of supply chain management, it is mainly limited to planning and transportation [9]. Therefore, how can the most effective and efficient algorithms be identified and utilized to improve the accuracy of predicting purchase lead times?

The paper is organized into four sections. The second section reviews previous works on Purchasing Lead Time Forecasting in the Industry. The third section presents a case study of a real-world company, while the fourth section summarizes this study's main findings and recommendations.

2. A Review of Previous Works on Purchasing Lead Time Forecasting in the Industry

This section examines previous works of purchasing lead time forecasting in the industry, focusing on essential techniques for improving prediction accuracy and managing inventory and procurement functions.

To gather relevant information, Google Scholar, Science Direct, IEEE, and SNDL were used as primary search tools, with the following keywords employed: "purchasing lead time," "order lead time," "forecasting techniques," "procurement and inventory," "industry 4.0," "machine learning," and "supplier delivery times."

Proceedings of M. F. E. G. I. M. - 2022/2023

six papers were selected out of thirty based on their relevance to our work scope. These papers present diverse techniques for predicting purchasing lead time and enable us to identify the most accurate method by comparing predicted results with actual data, calculating errors, and loss function.

1- In [8] the authors explored the Dynamic Forecasting of Lead-Time for Temporal Shipment Consolidation, which aims to help logistics companies optimize temporal shipment consolidation by proposing a dynamic lead-time forecasting method. The article suggests that the proposed machine learning-based approach can consider various factors such as supplier availability, production capacity, and shipping time. The article makes this prediction by using two algorithms. The first one is binary logistic regression. The second algorithm used is linear regression, a statistical method used to model the relationship between a dependent variable (lead time) and one or more independent variables (supplier availability, production capacity, shipping time, the size of demand, the number of unsatisfied orders, and the period since the Order was opened). The algorithm tries to fit a linear equation to the data that best represents the relationship between the variables [10]. The resulting model can predict the dependent variable's value based on the independent variable's values.
The study also discusses the value of freight consolidation to reduce CO2 emissions, but decision-makers must consider its disadvantages and the impact of increasing lead time. The paper also suggests supplier companies might provide better forecasting than third-party logistics companies. Finally, the study notes that temporal consolidation might not be attractive if inventory holding costs are high and transportation costs are low [8]. The article confessed that the results showed reasonable accuracy expressed in different ways, and as a final result, the article finds that the best-performed algorithm was the logistic regression with an accuracy of 93%. This means that among 100 predicted values, we can have only seven lead times incompatible with the real-world data, which will help the company consolidate. Finally, the study notes that temporal consolidation might not be attractive if inventory holding costs are high and transportation costs are low.

2- In [11] the authors stated that forecasting and purchasing lead time benefit the pharmaceutical sector. The article presents a study that applies five machine-learning regression algorithms to forecast purchasing lead times in a pharmaceutical supply chain using a real-world industrial database provided by Coopservice. The total number of samples was 42,753 collected during 2018, and it contains many features such as the day of the month of the customer order, weekday of the customer order, month of the customer order, supplier code identifier, product name identifier, and pharmaceutical product type category.
The algorithms are kNN, SVM, RF, LR, and MLP. The article compares the performance of these algorithms based on different metrics, such as mean absolute error MAE. It also discusses the advantages and limitations of each algorithm. The article mentions that most intelligent methods used in recent research do not use real-world data; instead, they create data through computer simulations and consider some simplifications for the internal manufacturing process. This work also clarifies that simulated databases frequently assume a flawless production system without factoring in machine failures, maintenance breaks, and delays in raw material deliveries. They performed two experiments to compare the machine learning systems employed in their analysis, splitting the whole corpus by category and Month. They used a standard 10-fold cross-validation procedure, dividing the entire dataset into ten distinct groups called folds, to choose the best hyperparameters for each algorithm. The article performed two separate experiments, and the main conclusion for the first experiment, which is the partition of data by category (Tumors, Diagnosis, Medicine ...), is that Linear regression outperforms the other algorithms in terms of accuracy and robustness, achieving the lowest MSE by 2.31 in average which is mean that if we use the model for predictions we will have an error of only two days, for the second experiment where the data is partitioned by months, The article found that the SVM gives best results by an average MSE equal to 1.89 which is mean that approximately can have two days only as gap with the real-data world. The article suggests that machine learning techniques can improve purchasing lead time forecasting in pharmaceutical supply chains and other industrial contexts.

3 - The article [12] the authors propose a data-driven approach to predict the deviation in supplier delivery performance from a buying company's point of view. The article claims that predicting deviation can help reduce the availability loss for customer orders and the costs for rush transportation caused by delivery delays

Proceedings of M. F. E. G. I. M. - 2022/2023

or uncertainties and focuses on the first two phases of a supply chain: supplier lead time from material suppliers and truck arrival time from logistics service providers (LSPs). The article aims to identify the factors that affect the deviation in these two phases and construct machine-learning models to predict the deviation based on historical data. The authors use a real-world dataset from a case company that produces industrial equipment, the Volvo Group Service Market Logistics. The dataset contains 1.3 million records of purchase orders with 19 features (Standard price, Delivery precision, Order hits, SEM result, Regions, Logistics result, Delivery method, QPM score, Country, Logistics Audit Result, Received Quantity, TB Actual Volume ...) and two target variables (deviation in supplier lead time and deviation in truck arrival time). The dataset is split into training (70%) and testing (30%) sets. To enhance the data quality and suitability for machine learning, the paper employs a variety of data preprocessing techniques, including outlier detection, missing value imputation, feature engineering, and feature selection to use a fair dataset that's why the number of features decreases to 14 for both RF catboost algorithms. The paper trains and tests two machine learning models: a classification model for predicting deviation in supplier lead time and a regression model for predicting deviation in truck arrival time. The article uses different SLT (Statistical Learning Theory) algorithms: 1 (see the definition in the appendix), 2 Gradient Boosting (see the definition in the appendix), 3 Random forests (see the definition in the appendix) The article uses different metrics, such as the F1-score, a measure of a model's accuracy that considers both precision and Recall (see more in the appendix), and the confusion matrix (see the definition in the appendix). As a result, the catboost outperforms the other two methods' recall scores for deviation prediction in the SLT model with values of 0.78; 0.63; 0.67 for the precision, Recall, and F1 scores, respectively. Random Forest has the highest recall score for the truck arrival time model, while Catboost performs better in precision. The average scores of each algorithm for the deviation class (early and late) are less than 0.5 in both models, which could consider this result so far from the deploy level, which should be better, at least above 0.8. The article also finds that some features such as order quantity, order value, supplier location, LSP type, and truck type significantly impact the deviation in supplier lead time and truck arrival time. (see the definition in the appendix 5), 2 Gradient Boosting (see the definition in the appendix5) , 3 Random forest (see in the appendix 5) The article uses different metrics, such as the F1-score, a measure of a model's accuracy that considers both precision and recall (see more in the appendix 5) and the confusion matrix (see the definition in the appendix 5). As a result, it finds that the catboost outperforms the other two methods in recall scores for deviation prediction in the SLT model with values of 0.78; 0.63; 0.67 for the precision, recall, and F1 score, respectively. For the truck arrival time model, Random Forest has the highest recall score while Catboost performs better in precision .The average scores of each algorithm for the deviation class (early and late) are less than 0.5 in both models which could consider this result so far from deploy level which should be better at least above 0.8 .The article also finds that some features such as order quantity, order value, supplier location, LSP type, and truck type significantly impact the deviation in supplier lead time and truck arrival time.

4 - In [13] the author present a case study where different machine learning algorithms are applied to predict the manufacturing lead time of semiconductor products. It claims accurate lead time prediction is crucial for improving production planning and scheduling, especially in complex and dynamic environments such as semiconductor manufacturing. The article aims to compare and evaluate the performance of 10 algorithms: 1. LR (Linear Regression) 2. Ridge and Lasso Regression 3. RT (Regression Trees) 4. Bagged RT (Bagged Regression Trees) 5. RF (Random Forest) 6. Boosted RT (Boosted Regression Trees) 7. SVM (Support Vector Machine) 8. MARS (Multivariate Adaptive Regression Splines) 9. kNN (k-Nearest Neighbors) 10. ANN (Artificial Neural Network) This article also discusses the challenges and benefits of using machine learning for lead time prediction, such as data quality, data preprocessing, feature selection, model training, model testing, and model interpretation [14]. The article uses a real-world dataset from a semiconductor manufacturer that contains 1.2 million records of production orders with 40 features and one target variable (lead time). The dataset is split into training (80%) and testing (20%) sets. The article applies various data preprocessing techniques such as outlier detection, missing value imputation, normalization, and feature engineering to improve the data quality and suitability for machine learning. As a result, it used just the ten most important variables of the models (1 Moving

Proceedings of M. F. E. G. I. M. - 2022/2023

average of the inter-departure times of the last 20 lot-layers; 2 Hours of the arrival time ; 3 Weekday of the arrival time; 4 Median of the product's lead time (in the analyzed process steps); 5 Work in progress: number of lot-layers in the analyzed process steps; 6 Work in progress: expected work content in minutes by process step bakefuse; 7 Capacity utilization of machines in process step sputter on the previous day; 8 Moving average of the inter-arrival times of the last ten lot-layers; 9 Mean of median operations times of a product on a given route; 10 Time in standby status of the machines of sputter on the previous day). In this article, ten algorithms are trained and tested using different combinations of features and hyperparameters. The article uses five metrics to measure the performance of the algorithms and evaluate the accuracy: mean absolute error (MAE), root mean squared error (RMSE), The normalized root mean squared error NRMSE, Mean square error MSE and Mean absolute percentage error (MAPE), it reports the best results for each algorithm based on these metrics and finds that the Random forest (RF) algorithm outperforms the other algorithms in terms of values by minutes of all this metrics by MAE 390; MAPE 33,8; MSE 360780; RMSE 600; NRMSE 12,5.

5 - In [15] the authors proposes a data-driven approach to predict the delivery times of purchase orders (POs) from suppliers to original equipment manufacturers (OEMs) in supply chain operations. The article claims that accurate prediction of PO delivery times is essential for improving production planning and scheduling, inventory management, and customer satisfaction. The authors aim to compare and evaluate the performance of supervised machine learning models in the form of Random Forests and Quantile Regression Forests trained on historical PO transactional data. The authors also discuss the challenges and benefits of using dimension reduction techniques such as principal component analysis (PCA) to reduce the number of features and avoid overfitting. The article uses a real-world dataset from an OEM that contains 1.4 million records of POs with 18 features such as (Item ID, Quantity Ordered, High-level part classification based on the functional requirement, Planned delivery time, Vendor ID, Material Group ID, Material Number ID, Plant ID...) and one target variable (delivery time). It applies various data preprocessing techniques such as outlier detection, missing value imputation, normalization, and feature selection to improve the data quality and suitability for regression models. It trains and tests the five regression models using different combinations of features and hyper parameters (Linear regression - Random Forest without Dimension Reduction - Random Forest with Dimension Reduction - Quantile Regression Forest without Dimension Reduction - Quantile Regression Forest with Dimension Reduction). The article reports the best results for each model based on these metrics and finds that the Random Forest model with PCA performs the best in terms of less error, and the best model improves the prediction accuracy by more than 30% for 94% lead time POs corresponding to unreliable supplier estimates. The article also finds that dimension reduction techniques can significantly improve the accuracy and efficiency of regression models (for the random forest decrease the error from 11.11 to 9.77 days and for Quantile regression from 11.22 to 9.90 days).

6 - In [1] the thesis focuses on defining and implementing a method for predicting the lead times of purchase orders using a gradient-boosting machine as the prediction model. The model uses purchase orders from the past to find patterns that can be used to estimate future lead times. The experiments were designed with three goals in mind. The first goal was to find the optimal hyper-parameters (cat_l2, cat_smooth, colsample_bytree, early_stopping_rounds, learning_rate,max_bin,min_child_samples,min_data _in_bin,min_data_per_group,min_gain_to_split, num_leaves, reg_alpha, reg_lambda, subsample, subsample_freq) for LightGBM (see the definition in the appendix 5). The second goal was to identify the features from the data with the best predictive power as there are more than ten features chosen that we can mention (Item group, Location of the storage, Supplier) as categorical choices, and (Estimate of the goods receipt processing time,Month when the PO was made,Year when the PO was made, Weekday of the supplier's delivery date estimate) as numerical features chosen. The first two goals were combined into one process using TPE (see definition in the appendix 5). The third goal was to produce a model with minimal complexity to save resources and increase generalizability. Unfortunately, the first and third goals contradict each other, so a reasonable balance had to be found. Three models were trained and tested using different combinations of features and hyper parameters. The article uses three metrics to measure the performance of the algorithms and evaluate the accuracy using MAE, MSE, and R-squared. As

Proceedings of M. F. E. G. I. M. - 2022/2023

a final result, according to the metrics cited, model one is more accurate by values of MAE, MSE, and R-squared (3.74; 130.8; 91.73) respectively, which means that the model in 100 predictions can find up to 91 right and the error will be estimated to average only 3.74 days , and according to the prediction time, model three is the most faster by a lower value this means that it doesn't take much time to give the lead time predictions and this helps the company to make decisions quickly.

Papers	Used algorithms	Limitations	Date of consultation of the citations
1	-binary LR -linear regression	- Accurate forecasting based on little information is challenging, indicating that there might be a lack of data for training the machine learning model. - The study assumes that the lead time is long, and accurate data about delivery times are provided by suppliers, which may not be applicable to all supply chain scenarios, highlighting the importance of understanding the assumptions and limitations of machine learning models. - The study used weighted average, linear regression, and logistic regression to perform the forecasting, indicating that choosing the appropriate machine learning model for the problem at hand can be challenging.	April 2023
2	- SVM - RF - LR - MLP	- The forecast of lead time for purchasing products alone may not be enough to provide decision-makers with a comprehensive view of the entire supply chain lead time, leading to potential bottlenecks or disruptions in the supply chain. - The current methods for forecasting internal supply chain lead times for real service processes may not be accurate or comprehensive, leading to inefficiencies and delays in the supply chain.	April 2023
3	- cat boost - Gradient descent - RF	- There are outliers in historical deviation performance that are not examined, but they are kept in the dataset and contribute to the machine learning of minority classes. This could affect the model's performance. - Missing values in the features are handled using average values, which may result in a poor estimation of variances and correlations for the feature. Advanced imputation methods remain to be investigated.	April 2023

Proceedings of M. F. I. G. I. M. - 2022/2023

4	- LR - RLR - RT - Bagged RT - RF - Boosted RT - SVM - MARS - KNN - ANN	- The presence of interrelations between variables with negative sensitivity may impact the accuracy and effectiveness of the analysis, underscoring the need for further research to better understand and address these relationships.	April 2023
5	- Linear regression - Random Forest - Quantile RF	- The time required to train regression models and dimension reduction methods for large datasets may be a limiting factor in achieving real-time predictions based on continually updated PO delivery data, highlighting the need for parallel implementation to expedite the process. - The current use of median quantiles in the Quantile Regression Forest model may not be the most effective approach for estimation, suggesting the need for further investigation into the performance of different quantiles to improve prediction accuracy.	April 2023
6	- Light GBM by TPE	- The lack of information about the reliability of model predictions may limit their usefulness, highlighting the need for the development of prediction intervals to evaluate the uncertainty of point predictions and perform further analysis to understand the reasons behind inaccuracies. - The current approach for training GBMs using quantile regression as a cost function may not be the most effective approach for generating prediction intervals, suggesting the need for further optimization of hyperparameters using alternative cost functions	April 2023

Table 1: Summary of the selected papers

Proceedings of M. E. E. G. I. M. - 2022/2023

According to the analysis of the selected articles table 1, we can notice that the methods include Linear regression [8, 11], KNN, SVM, RF and MLP [11, 12], Gradient Boosting, CatBoost [12]. The articles above mention the challenges that can affect the model such as the data used, and we find that most have used real-world data. We take as an example the article [13] and [15] both used real-world data that contained millions of records that allowed them to have accurate results ;cause the problem in the use of simulated data is well clarified in [12] of which mentions that simulated databases frequently assume a flawless production system without factoring in machine failures, maintenance breaks, and delays in raw material deliveries. As it has been said that the quality of the data is very important for a valid study and that's why the papers employ a variety of data preprocessing techniques, including outlier detection, missing value imputation, feature engineering, and feature selection to use a fair dataset [1] which also helps to reduce and eliminate and keep only the valuable elements in our data [15]. Furthermore some authors find that dimension reduction techniques can significantly improve the accuracy and efficiency of regression models to reduce

the number of features and avoid overfitting [1]. For the data used in the predictive model generally, we find that it's divided into two parts: 80% for the training and 20% for the test, atypically for [12] of which it changed the portion 70% for the training and 30% for the test which can give less accurate results. One of the most important factors in making a successful model is the choice of hyperparameters, and there are several methods to choose them. [11] used a standard 10-fold cross-validation procedure, dividing the entire dataset into ten distinct groups called folds, to choose the best hyperparameters for each algorithm. [1] used a Hyper-parameter optimisation via tree-based Parzen estimator TPE. To evaluate all these experiences and the different models, there are some very relevant and significant metrics such as MSE [11, 13], recall, and F1 score [12], MAE [13] and R-squared [1] . Others are rarely used such as, MAPE, RMSE and NRMSE [13].

The experiments made in those studies cannot be directly compared because of a different domain, dataset, and features, and the major difference between them is that they have multiple steps where the person applying the method has to make choices dependent on the dataset. The choices include the number of categories to keep before and after PCA, the number of principal components to consider, and the hyper-parameters depending on the algorithms used, but according to the analysis of each one. As a result, we can notice that among the algorithms which gave good accuracy in the problem of forecasting purchasing lead time, we quote: The algorithm that has proven its effectiveness is the random forest, which was found to be the best algorithm in three different experiments due to his capability to handle high-dimensional feature spaces. [12, 13, 15]., The standard Linear regression is a simple and easy algorithm to understand. , SVM, which is Robust to noise and outliers in the dataset of the pharmaceutical case study [11], the catboost also one of the more accurate algorithm with good performance on large and complex datasets and can treat the data quickly [12] .

3. Case study

This case study uses machine learning techniques to predict supply lead times in an Algerian oil services company. The company is responsible for developing, producing, transporting, and marketing hydrocarbons in Algeria.

3.1. Methodology and Dataset description

Figure 1 provides an overview of the machine learning process's different phases. The first phase is preparing the database. This study used a database of 20,989 samples collected over five years. These samples were carefully selected to form the training, validation, and test sets. The independent variables, consisting of 22 categorical and numerical variables, were considered input vector x for the machine learning systems to predict the delays. These lead times were derived from the company's enterprise resource planning (ERP) system and are presented in Tables 2 and 3. There are different methods to select the most relevant features in a dataset. One of these methods is Recursive Feature Elimination (RFE), which involves training a model with all features and iteratively removing the least important ones until a predefined threshold is reached. This method uses a feature selection algorithm that evaluates the importance of each feature by removing them iteratively [16] [17] . Similarly, the Stepwise Regression method involves iteratively selecting features by adding or removing one variable at each step until an optimal model is achieved [18][19] In our study, we used the Kruskal-Wallis ANOVA statistical test to assess the effect of a categorical variable on a continuous numerical variable [16]. This test is appropriate when the data distribution is abnormal, or the variance is not homogeneous between groups. It uses the ranks of the data to ignore their distribution [18]. The steps of the Kruskal Wallis test consist of ranking the data, assigning a rank to each data point, summing the ranks of the groups for each point, and calculating the statistic H. However, there is no unique formula to determine the effect of a categorical variable on a continuous numerical variable in a Kruskal-Wallis ANOVA. Therefore, some researchers have proposed formulas based on the calculated H statistic [20]. Samples were selected after exploratory data visualization, which was performed to detect possible anomalies. First, duplicates and corrupted data were removed to ensure data quality. Next, outliers were identified using box plots, which visualize the data graphically and identify values that deviate from the norm. These outliers were then removed to prevent them from affecting the quality of the analysis. Finally, missing values were handled appropriately to ensure data consistency. These preprocessing steps are essential to ensure that the data are ready for analysis and that the results are reliable.

Proceedings of M. E. E. G. I. M. - 2022/2023

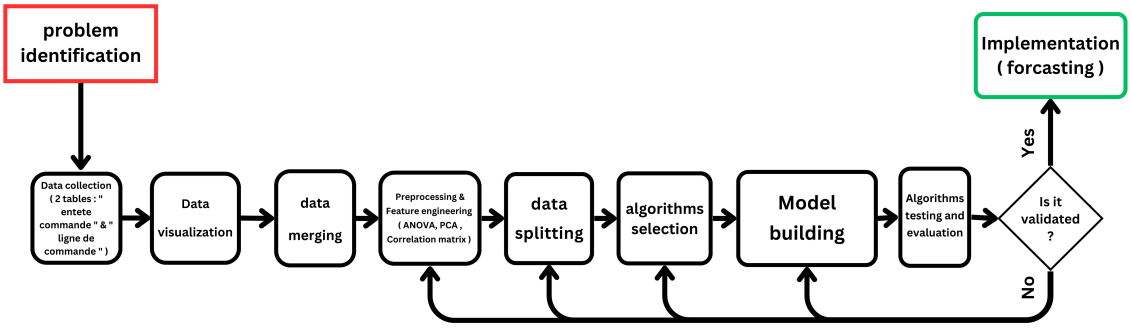


Figure 1: Overview of machine learning process

Feature name	Description
AOF_OFF_NUM	Offer number
AOF_OFF_NUM	Measurement Unit
AOF_OFF_NUM	Quantity assigned from the order item to the requests
LCMD_UNITE	Supplier code
LCMD_QUANTITE	Contract number
FOU_CODE	Agent in charge of the follow-up of the case
CTR_NUM	Regulation of the order
CMD_SUIVI	Method and place of delivery
CMD_REGLEMENT	Currency code
CMD_MODLIEULIV	Payment type code
CMD_MONNAIE	Cancel an Order

Table 2: Categorical features

Proceedings of M. G. I. M. - 2022/2023

Feature name	Description
LOT_NUM	Lot number
LCMD_PRIX_UNIT	Unit price of the order item
CMD_DATE	Order date
LCMD_PRIX_UNIT	Lead time of purchase
CMD_DATE	Packaging cost
CMD_DELAIACHAT	transport cost
CMD_EMBALLAGE	AVENANT number
CMD_TRANSPORT	Delivery time per order
CMD_AVENANT	Direct Purchase
CMD_DELAICMD	Apply the discount on Transport Cost
CMD_ADIRECT	Apply discount on Packing cost

Table 3: Numerical features

After feature selection and data preparation, we conducted a Pearson correlation coefficient analysis (see in figure 2) for each numeric variable against the target CMD_DELAIACHAT. This analysis assessed the strength and direction of the linear relationship between these two continuous variables. The results showed that most variables did not correlate significantly, except for CMD_DATE, which had a moderate negative correlation. This correlation suggests an inverse relationship between order dates and purchasing lead times. The farther the ordering date is, the shorter the purchasing lead times. This observation is essential for order planning and inventory management, as an order placed too early can lead to an increase in inventory, while an order placed too late can cause a stock out. A similar trend is observed for direct purchasing, with a weak negative statistical correlation. This means the direct purchase probability decreases slightly as the purchase time increases.

3.2. Experimental results

We started by standardizing the data in our comparative study of machine learning systems. Standardization is a common data preprocessing technique that involves centering the data around zero and scaling it to have a unit variance. This is done by subtracting the mean of each numerical variable and dividing the result by the standard deviation of that variable, thus putting all variables on the same scale and making it easier to compare. We also transformed the data by applying a mathematical function to convert them into a new form. This transformation was performed to normalize the data's distribution or reduce its variance. Next, we used principal component analysis (PCA) to reduce the complexity of several machine learning models (see in table 4) , including KNN, DECISION TREE, SVR, CATBOOST, RANDOM FOREST, MLP, and DNN. For this reduction, we set the number of principal components to 10, except for the LGBM model, for which we put two components because of the complexity of its training data. We used 80% of the data set to train all models. Finally, we evaluated the performance of the models by comparing their predictions with the observed lags in the test set, using three different error measures: the MAE (mean absolute error), the MSE (mean square error), and the coefficient of determination (R2) (see in figure 3) Our experimental design aimed to find the optimal hyperparameters for each model. To accomplish this, we used the RandomizedSearchCV method, which randomly selects the hyperparameter values to be tested within a defined range for each hyperparameter, and then tests each random combination of hyperparameter values to find the best combination that works for the chosen data set and model. The results of hyperparameters are shown in table 4 which summarizes the results obtained for each model with the optimal hyperparameters. Finally,Figure 3 presents the final result of our comparative study.

Proceedings of M. F. E. G. I. M. - 2022/2023

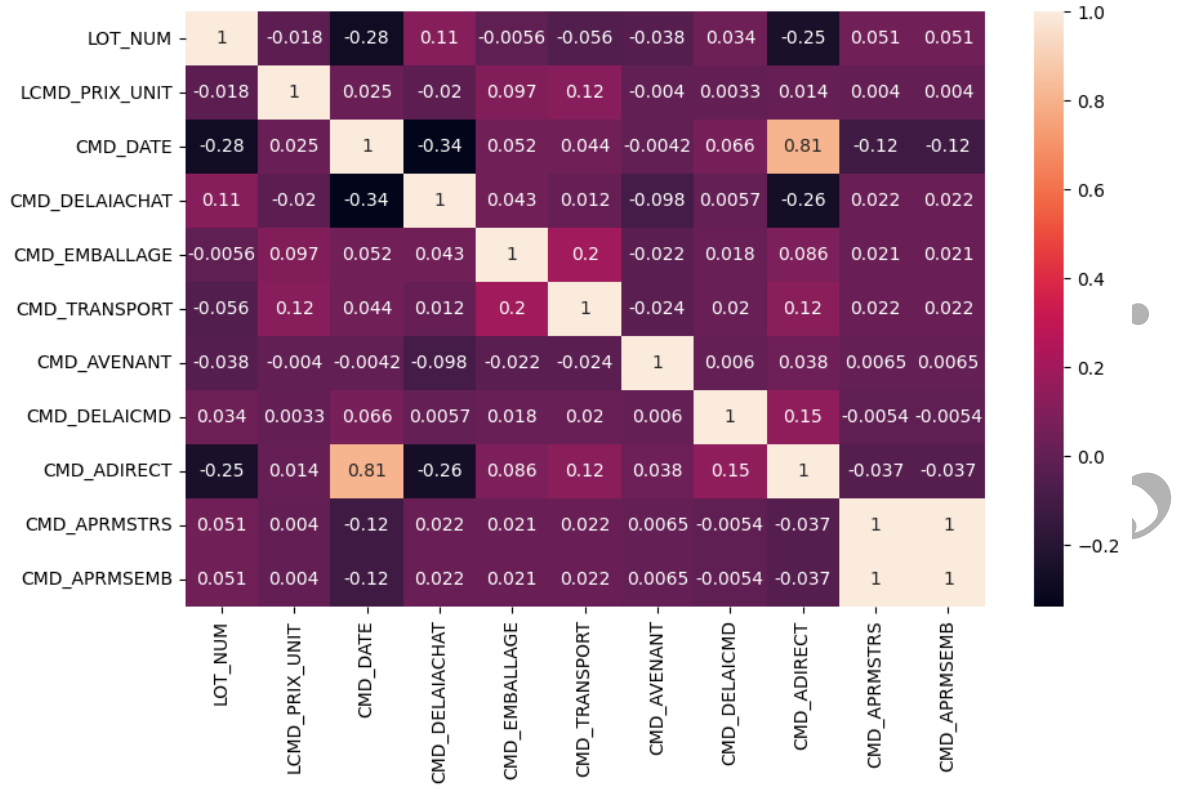


Figure 2: Pearson correlation coefficients for the numerical features and the target variable CMD_DELAICMD

Models	hyper parameters
KNN	weights='distance', p=1, n_neighbors=3
LGBM	learning_rate=0.5,max_depth=5,min_child_samples=16, n_estimators=1230, num_leaves=9, reg_alpha=0.01, reg_lambda=2.0
Decision Tree Regressor	max_depth= 19, min_impurity_decrease=0.0001, min_samples_leaf=9, min_samples_split=3
Cat boost regressor	max_depth=16, iterations=100, learning_rate=0.1
MLP	hidden_layer_sizes=(88,50) ,solver='adam',max_iter=300,random_state=42, activation='tanh'
DNN	Input_layers=109,Hidden_layers=57,output_layer=1,activation='tanh', optimizer='rmsprop',epochs=210,batch_size=55
SVR	kernel='rbf', C=186.3133, gamma=1
Random Forest Regressor	n_estimators=882,max_depth=18, random_state=42,min_samples_split=17, min_samples_leaf=10

Table 4: Model hyper parameters

.

Proceedings of C. I. M. - 2024

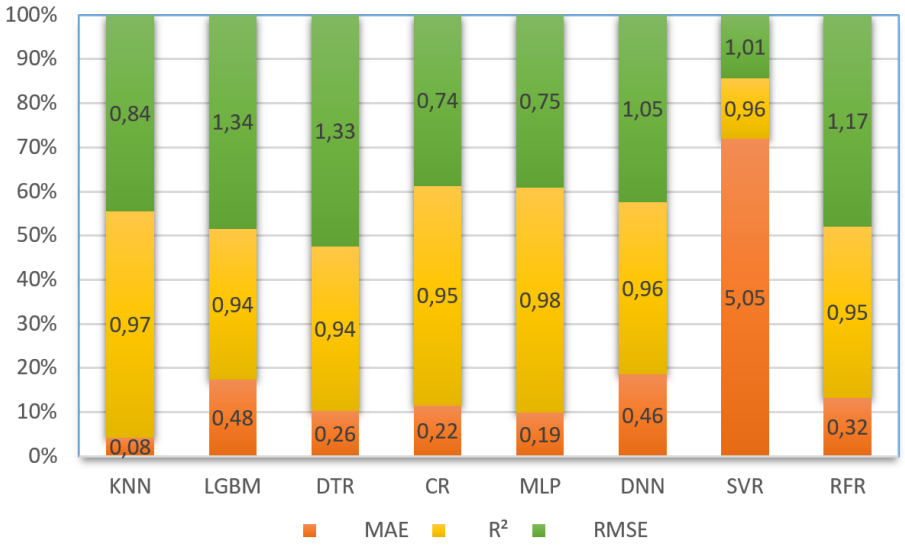


Figure 3: Error measures for each model

The following figure 3 presents error metrics for the different models used in this case study. The coefficient of determination R^2 is used to evaluate the quality of the models' fit to the data. A high R^2 indicates a good fit, while a low R^2 indicates a poor fit. The RMSE and MAE metrics are also used to evaluate the accuracy of the model's predictions. RMSE is more sensitive to outliers, while MAE is more robust [19]. The model with the lowest MAE or RMSE value is considered the best to achieve accurate predictions. However, it is essential to note that some models may perform well but suffer from overfitting, meaning they have memorized the training data rather than generalizing it to new data. Overfitting can occur when the model is too complex relative to the quantity and quality of the training data or too trained. To mitigate this issue, regularization was added to the models to decrease the importance of the cost function weights, which favors smaller and more regular coefficients.

4. Conclusions

This paper presents the steps and methods to be applied to find the best machine-learning algorithm for forecasting lead times in the supply chain. Indeed, we started with the preparation of the data, as it is crucial and requires a thorough analysis, then the selection of the most relevant features by using statistical methods and the visualization of correlations to assess the impact of categorical and numerical variables on the target variable. Eliminating outliers and the dimensionality reduction method are also essential to obtain accurate data. The second phase consists of applying different models using hyperparameter optimization methods such as Bayesian random search and cross-validation and regularization to decrease the weights of the cost functions of the models to avoid overlearning. Experimental results show that it is possible to predict the purchase delay with high accuracy, especially for the models KNN, MLP, CATBOOST, and RANDOM FOREST, unlike SVR, knowing that it should be noted that the basic algorithms such as KNN and DT may suffer from overlearning or not fit the data perfectly. Besides, the hyperparameter selection methodology implemented in the presented study is limited by time and results, making it difficult to generate these hyperparameters for other situations. To further improve the accuracy of the results, it would make sense to leverage additional data, including production and sales data, instead of focusing exclusively on supply data. To address complexity issues, it is recommended to use deep learning models and more advanced hyperparameter optimization methods, such as genetic algorithms and reinforcement learning.

References

[1] Juho-Pekka Koponen. Predicting lead times of purchase orders using gradient boosting machine. 2020.

[2] Augusto Bianchini, Andrea Benci, Marco Pelle-

Proceedings of M. C. I. M. - 2022 / 2023

grini, and Jessica Rossi. Supply chain redesign for lead-time reduction through kraljic purchasing portfolio and ahp integration. *Benchmarking: An International Journal*, 26(4):1194–1209, 2019.

[3] Jin-Shan Yang and Jason Chao-Hsien Pan. Just-in-time purchasing: an integrated inventory model involving deterministic variable lead time and quality improvement investment. *International Journal of Production Research*, 42(5):853–863, 2004.

[4] Yanhua Xiong. Coordinating order decisions for assemble-to-order enterprise with uncertainty in purchase lead time and demand order. In *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, pages 4591–4596. IEEE, 2011.

[5] Katherine L Milkman, Todd Rogers, and Max H Bazerman. I'll have the ice cream soon and the vegetables later: A study of online grocery purchases and order lead time. *Marketing Letters*, 21:17–35, 2010.

[6] Hardik Meisheri, Nazneen N Sultana, Mayank Baranwal, Vinita Baniwal, Somjit Nath, Satyam Verma, Balaraman Ravindran, and Harshad Khadilkar. Scalable multi-product inventory control with lead time constraints using reinforcement learning. *Neural Computing and Applications*, 34(3):1735–1757, 2022.

[7] Kut C So and Xiaona Zheng. Impact of supplier's lead time and forecast demand updating on retailer's order quantity variability in a two-level supply chain. *International Journal of Production Economics*, 86(2):169–179, 2003.

[8] Mohammed Alnahhal, Diane Ahrens, and Bashir Salah. Dynamic lead-time forecasting using machine learning in a make-to-order supply chain. *Applied Sciences*, 11(21):10105, 2021.

[9] Halima Bousqaoui, Said Achchab, and Kawtar Tikito. Machine learning applications in supply chains: An emphasis on neural network applications. In *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, pages 1–7. IEEE, 2017.

[10] 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.

[11] Maiza Biazon de Oliveira, Giorgio Zucchi, Marco Lippi, Douglas Farias Cordeiro, Núbia Rosa da Silva, Manuel Iori, et al. Lead time forecasting with machine learning techniques for a pharmaceutical supply chain. In *ICEIS (1)*, pages 634–641, 2021.

[12] Meng Huang and Masood Bagheri. Predicting deviation in supplier lead time and truck arrival time using machine learning-a data mining project at volvo group. 2019.

[13] Lukas Lingitz, Viola Gallina, Fazel Ansari, Dávid Gyulai, András Pfeiffer, Wilfried Sihn, and László Monostori. Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. *Procedia Cirp*, 72:1051–1056, 2018.

[14] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

[15] Jundi Liu, Steven Hwang, Walter Yund, Linda Ng Boyle, and Ashis G Banerjee. Predicting purchase orders delivery times using regression models with dimension reduction. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 51739, page V01BT02A034. American Society of Mechanical Engineers, 2018.

[16] StatSoft. Test de kruskal-wallis, 2021.

[17] Fabian Pedregosa and Varoquaux. Scikit-learn documentation: Recursive feature elimination. `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html`, 2021. [Online; accessed 10-May-2023].

[18] Maciej Tomczak and Ewa Tomczak. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in sport sciences*, 21(1), 2014.

[19] Ryan Kwok. Stepwise regression tutorial in python, 2023. Accessed: 08-05-2023.

[20] Statistics Formula Glossary. University at buffalo, 2019. Accessed: 08-05-2023.

Proceedings of M. F. E. G. I. M. - 2022/2023

5. Appendix

Concept	Definition
Forecasting	Part of the prediction involves estimating future events using historical data and adding an explicit time dimension
CatBoost	CatBoost is a gradient boosting algorithm that Yandex developed. It is designed to handle categorical features in data sets and incorporates several advanced features, such as ordered boosting, which enables the algorithm to consider the Order of instances during training. It also includes a built-in method for handling missing data.
Gradient Boosting	Gradient boosting is a machine learning technique for building predictive models, typically for regression and classification tasks. It works by iteratively training decision trees on the residuals of the previous iteration, where the residual is the difference between the actual and predicted values. In each iteration, the model tries to find the optimal tree that minimizes the loss function, which is typically a measure of the difference between the predicted and actual values.
F1-score	It is the harmonic mean of precision and recall, where precision is the proportion of true positives among all the predicted positives, and recall is the proportion of true positives among all the actual positives. The F1 score is commonly used in binary classification tasks, where there are positive and negative classes. The F1 score ranges from 0 to 1, where one indicates perfect precision and recall, and 0 indicates poor performance)
Confusion Matrix	The confusion matrix is a performance measurement technique for a machine learning classification algorithm. It is often used to measure the performance of classification models), Precision, and Recall to measure the performance of the models.
LightGBM	LightGBM is a gradient-boosting framework that uses decision trees to perform the classification or regression tasks. It is similar to other gradient boosting frameworks but has some unique features, such as its ability to handle large datasets using histogram-based algorithms for computing gradients efficiently. LightGBM also uses leaf-wise growth instead of level-wise growth, which can result in faster convergence and higher accuracy to accurately predict purchase order lead times.
TPE	TPE stands for Tree-structured Parzen Estimator. It is a Bayesian optimization algorithm that automatically optimizes hyper parameters 1. TPE has been used to simplify machine learning for tasks such as genomic prediction by optimizing the hyper parameters of machine learning methods like Support Vector Regression (SVR)
MARS	MARS (Multivariate Adaptive Regression Splines): MARS is a regression algorithm that uses a collection of linear regression models to fit the data. It partitions the data into smaller subspaces and fits a linear regression model in each subspace.

Proceedings of M. F. E. G. I. M. - 2022/2023

KNN	**(k-Nearest Neighbors): kNN is a non-parametric classification algorithm that assigns the class of a new data point based on the majority class of its k-nearest neighbors in the training data [0],**
SVM	**support vector machines SVM (A supervised machine learning algorithm that can be used for both classification and regression tasks. It works by finding the hyperplane that best separates the data points into their respective classes, with the maximum margin of separation between the classes) [0].**
RF	**Random forests (a supervised machine learning algorithm used for classification, regression, and other tasks. It is an ensemble method that builds multiple decision trees and combines their results to make a final prediction.)**
LR	**linear regression (a supervised machine learning algorithm used for regression tasks. It works by finding the line of best fit that minimizes the distance between the predicted values and the actual values in the training data)**
MLP	**A multilayer perceptron's MLP is a type of artificial neural network that is used for a variety of supervised learning tasks, including classification and regression. It consists of multiple layers of interconnected nodes (or neurons) and is trained using backpropagation to minimize the difference between predicted and actual outputs.)**
RL	**Ridge and Lasso Regression: Ridge and Lasso are regularization techniques used in linear regression to prevent overfitting. Ridge regression adds a penalty term to the sum of squared errors, while Lasso regression adds a penalty term based on the absolute value of the coefficients.**
RT	**RT (Regression Trees): A decision tree algorithm is used for regression problems. It splits the data into smaller subsets based on the value of a selected feature and recursively creates branches until the data is adequately described.**
Bagged RT	**Bagged RT (Bagged Regression Trees): Bagging is an ensemble learning technique that involves building multiple models from bootstrap samples of the training data. Bagged Regression Trees use this technique to build multiple decision trees and combine their predictions to reduce variance and improve accuracy.**
ANN	**ANN (Artificial Neural Network): An artificial neural network is a machine learning algorithm modeled after the human brain's structure and function. It consists of layers of interconnected nodes that perform mathematical operations on the input data to learn and make predictions.**

Proceedings of M. F. E. G. I. M. - 2022/2023

binary logistic regression	**Binary logistic regression is a statistical method used to model the probability of a binary outcome 0 or 1 (where 1 means delivery of a specific order is expected to occur, and 0 means delivery does not occur) based on one or more predictor variables. The algorithm calculates the odds of the binary outcome and converts them into a probability value between 0 and 1 ,**
DNN	**DNN (Deep Neural Network) is an artificial neural network that consists of multiple layers of interconnected nodes or neurons. It is used in machine learning and AI applications to process input data and generate output results. DNNs are "deep" because they have many hidden layers that allow them to learn complex patterns and relationships in the data. They are trained by adjusting the weights of the connections between neurons through a process called backpropagation.**
ANOVA	**A statistical test is used to determine the effect of a categorical variable on a continuous numerical variable. This test is appropriate when the normal distribution of the data and residuals is not respected, or when the variance is not homogeneous between the groups being compared (heteroscedasticity of the data).**

Proceedings of M. F. E. G. I. M. - 2022/2023