Use of a Machine Learning Model for the Reduction of Backorders in the Cross Docking Sales Process for the Homecenter Order Service

Garcia Lopez Yvan Jesus
Senior Lecturer
Faculty of Engineering and Architecture
Industrial Engineering Career
Lima University
Lima, Peru
orcid.org/0000-0001-9577-4188
ygarcia@ulima.edu.pe

Jamil Panduro y Sebastian Pumayauri
Faculty of Engineering and Architecture
Industrial Engineering Career
Lima University
Lima Peru
orcid.org:0000-0002-1303-6200/ 0000-0001-7790-8429
20173692@aloe.ulima.edu.pe , 20173749@aloe.ulima.edu.pe

Abstract

In this work, it is necessary to analyze the increase of Back Order in the attention of crossdocking orders in the attention of Homecenter customers due to the lack of definition of purchase planning processes, resulting in logistics costs, fill rate charges and low service level. Thus, it is intended the companies that handle high volumes of inventory and constant orders should have a forecast plan to cover possible stock-outs. The main purpose of the research is to explain a way to prevent stock-outs using an artificial intelligence model, based on historical sales data of a medium-sized company that manages inventories, as well as to determine the machine earning model to predict and reduce backorders. For the data analysis, the Orange software was used, where the data was trained with different artificial intelligence models such as Decision Tree, Support Vector Machine, Random Forest, and neural networks. The most accurate model was defined according to numerical indicators such as the confusion matrix, the area under the curve (AUC) and the ROC curve analysis. Thus, we opted for the neural network model, which presented the most accurate data. Finally, the results are presented and a suggestion is made at the management level regarding decision making in the supply process. For this purpose, itis considered pertinent to delve into the subject of the variables that influence the accumulation of backorders.

Keywords
Machine learning, Backorders, Demand forecasting, Supply chain and inventory management.

1. Introduction
With the advent of the internet at the end of the 20th century, the information available to prepare highly effective mathematical models was massified (Rudin 2019). Also, the computing power of computers increased benefiting the automation of processes in different industrial and technological companies (Rudin 2019).

In the technology industry, the concept of "machine learning" is applied, a branch of artificial intelligence (AI), which allows the machine learning without the need to previously program them for it (Adadi y Berrada 2018). Currently, this indispensable ability allows systematization to identify patterns among data and predict them (Adadi y Berrada 2018)

Evidencing the technology-based proposal by this industry, the need for process automation is latent, and therefore, emphasis is placed on addressing this issue through models that reduce pending orders in "cross docking" sales processes.

On the other hand, it is also mentioned the improvement of attention of the "Home center" orders, whose relevance tour around the process logistics that speeds up the delivery of said orders. it is so as the problem of this research topic will focus _ in the application of "machine learning" for inventory management and effective order distribution.

1.1 Objectives

- The objective of the research is to determine the machine learning model to predict and reduce BackOrders.
- Identify the logistical variables with the greatest influence on the accumulation of the backorder
- Outline the model with mathematical algorithms with the variables to be analyzed

2. Literature review

The backorder is defined as that product within a purchase order that cannot be served due to lack of stock, ie. The product is not available due to different internal and external logistical factors, thus generating orders on hold for the customer, which in some cases is penalized (Islam and Amin 2020). On the other hand, cross-docking is a logistics technique that is based on reducing inventory costs to increase the quantity of the flow of goods or products, so as to improve the efficiency of the supply chain (Fonseca, et al, 2019). Scheduling decisions are quite relevant to ensure acceleration in inventory turnover and normalize on-time deliveries (Fonseca, et al, 2019).

The definition of the set of instructions defined by the reinforced learning models differs in being classification and regression (Higa 2021). The former has the disadvantage of not being suitable for large datasets due to its high training time as well as it uses two-dimensional data for its classification in the hyperplane, the latter, serves to predict results in quantity according to the value of the inputs or analysis variables. (Higa 2021).

Neural networks is one of the most complex artificial intelligence processes that requires greater processing capacity for collecting data and converting them into useful information in a shorter time, thus being a tool that helps in prediction and decision making (Papernot et al., 2017).

It consists of relating input variables by assigning a weight that defines the intensity of each of the input variables that, affects the neuron, which is the base processing unit within a neural network, where it is analyzed and processed to finally have output values or result. (Butler et al., 2018).

Decision tree is a tool that helps to forecast possible outcomes of a variety of decisions that through a set of data the algorithm has the autonomy to select a variable that generates a subdivision, and so on (Namazkhan, et al., 2019)

The authors Islam and Amin (2020). They proposed the predictive analysis of backorder using the 5-level decision tree model, where the logistic variables involved in the process were proposed. For this purpose, the relationship between the nodes was sought according to the correlation existing between the dependent and independent variables. For this, purpose, the Spearman correlation model was used for grouping the variables.

Support Vector Machine is the supervised learning Machine Learning algorithm of classification and regression prediction that analyzes binary variables in a two-dimensional plane, searches for the maximum separation between observations (Zhang, et al,2021). This algorithm is suitable for classifying a small to medium data set that has high level of complexity (Khan, etal, 2021). SVM aims to draw a hyperplane in an "n" dimensional vector space to separate it into two distinct data patterns representing the respective classes.

Zhang et al (2021), applied the vector support machine (VSM) model for the reduction of the Bullwhip Effect (demand estimation inaccuracy phenomenon) for risk minimization through the training of a series of observations in order to classify the result according to the assigned variables. For this reason, the sample or data processed should be as large as possible so that the algorithm can generate a stable result.

Random forest is an ensemble Machine Leaming technique whose adaptability to data is quite high, with the ability to
find correlation and interactions between variables (Lundberg et al. 2020). Models based on random forest can be more
assertive than neural networks in cases where the models have tabular style data (Lundberg et al. 2020). That is, when variables are individually significant and lack temporal or spatial, multiscale or structured data (Lundberg et al. 2020).

One of the processes that require more attention for any company in the world is inventory management, whose effectiveness is what will allow it to promote itself in more competitive business (Moshtagh and Taleizadeh 2017). So, itis necessary to predict backordered products before the customer places the order, production should be regulated to reduce lead times and increase profitability (Hajek and Abedin 2020)

Over the years, the industry dedicated to the distribution of products has been updated to be able to implement their mathematical models with machine learning tools to automate inventory management (Fonseca et al. 2019). Optimization models were developed to obtain better inventory policies with genetic algorithms and big data analysis, demonstrating the effectiveness of prediction (Hajek and Abedin 2020)

It should be noted that in this industry the concept of crossdocking has given good results in reducing the amount of stock in storage (Fonseca et al. 2019). With this, an exaggerated amount of money tied up in inventory was avoided: however, it has not been enough to meet the demands of customers regarding their orders (Fonseca et al. 2019). Most, customers do not consider companies with a long history of backorders as a good option despite the small amount of stock resulting in lost sales (Chan S. et al. 2017).

3. Method

The present research work has a quantitative approach and explanatory-correlational level. According to the correlational sequence is longitudinal, since, historical data of logistic variables (Forecasts, Sales, Inventory level, Backorder) between the years 2018 to 2020 are collected to be analyzed, the longitudinal sequence decomposes data from a range of time that are going to be studied (Manterola et al. 2019)

In addition, the methodology offers a final product of scientific explanation, due to the type of probability sampling, using statistical techniques for the analysis of the behavior of the variables and their main dimensions, in addition, Salgado (2019) mentions that, probability sampling analyzes a part of the population that meets the probability of being chosen at random and is divided into three types, stratified, simple and systematic.

The sampling technique used is stratified, since the universe of products will be divided into families according to their type of use. Stratified sampling separates or subdivides the population into groups of similar characteristics (Salgado 2019).

Next, the scope of the research is explanatory because it details the artificial intelligence tools used and the structure in which the algorithms are composed. Explanatory research is conducted for a problem that was not well investigated before, demands priorities, generates, operational definitions, and provides a better investigated model. In fact, itis a type of research design that focuses on explaining aspects of your study (Manterola et al. 2019). For this purpose, historical data was collected on transit time, current inventory level, sales and sales forecast, calculation, which are called numerical variables, followed by the identification of categorical variables that are highly relevant in the backorder prediction
.

4. Data Collection

Table 1. Definition of logistic variables

Item	**variables**	**Definition**
1	Demand	It is recorded from sales between the years 2018-2021
2	Demand projection	A projection of the demand of 3,6,9 months is made

3	products in poor condition	They are products that are found with some damage, and are stored for their return
4	transit time	It is the time in which the product enters the warehouse since the purchase order is issued.
5	Quantity in transit	It is the quantity that is on the way, either by sea or by air.
6	back -order	These are the orders that were not fulfilled due to lack of stock

The artificial intelligence model applied to the reduction of _ orders slopes search predict through the analysis of logistic variables identified , applying algorithms mathematicians who define the degree or intensity of impact of the variables in the forecast of the orders .

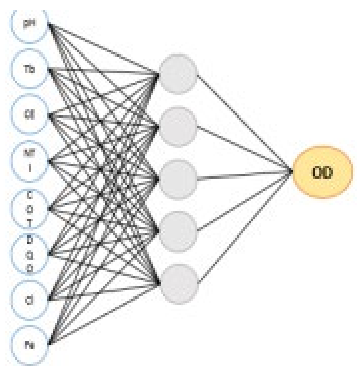

Figure 1. Design of a neural network

The artificial intelligence model is based on multiple regression that relates the dependent variables and a dependent variable (Figure 1-3), for an optimal predictive analysis a high multiple regression coefficient and low standard deviation are required (Marroquin et al. 2021)

Likewise, the model is based on relating the neural network, base input unit between the connections, assigning a weight (Constant) to each variable according to the multiple regression obtained previously, which serves to define the degree of intensity that each variable affects the neuron, in order to obtain a much more accurate result (Marroquia et al., 2021)

The use of neural networks for backorder calculation follows a line of forecasting based on the analysis of historical and future data to predict unusual events with respect to changing demand (Tang and Ge 2021).

Y=W1X1+W2X 2+W3X3....b

For this purpose, the mathematical model to be used is defined, where y is the dependent variable; n is the sample size ate the variables; W is the numerical constants obtained from the multiple regression and b is the bias that provides an orientation or direction to the analysis as required by the researcher.

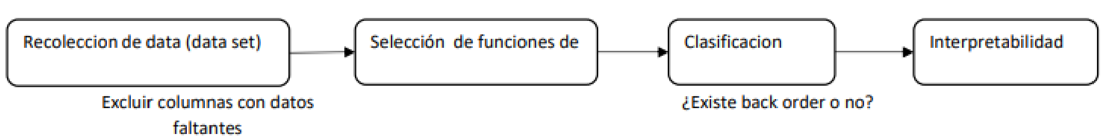

Figure 2 *Sequence of data analysis*

the process consists in collect the necessary data according to the variables to be a studied, then assigning a weight to each of them according to the degree of relevance or affect in the backorder, then the classifying the binary variables in order to have a more accurate interpretation.

5. Results and Discussion

5.1 Numerical Results

The results were made evaluation of different machine learning models was carried out, presenting indicators of the adjustment and hyper parameters performed to the models in the orange software (Table 1-2).

Table 2. Precision indicators of the artificial intelligence models _

ML model	Precision	AUC	Memory	confusion matrix	Hyper parameters
SMV	0.797	0.704	0.514	Nope Yes Nope 1000 1266 Yes 86 430	C=50; Kernel =Linear
random forest	0.964	0.967	0.964	No Yes Nope 2253 13 Yes 88 428	Number of trees =12
Decisions Tree	0.961	0.864	0.959	No Yes Nope 2266 0 Yes 115 401	No. _ Min instances on leaves = 2; Maximum tree depth= 100
neural network	0.995	0.994	0.995	No Yes Nope 2264 two Yes eleven 505	activation = tanh ; alpha = 0.0001, Solver = L-BFGS-B

Table 2 shows the efficiency evaluation indicators for the classification algorithms used for the analysis of this study.

According to the confusion matrix, itis determined that the neural network is the model that has classified the categorical variable more accurately, since the metric intuits the non-existence of back order in 2271 real data, the existence of back order in 503 data and the discrepancy of 12 data between the real and the prediction of the model is detected.

5.2 Graphical Results

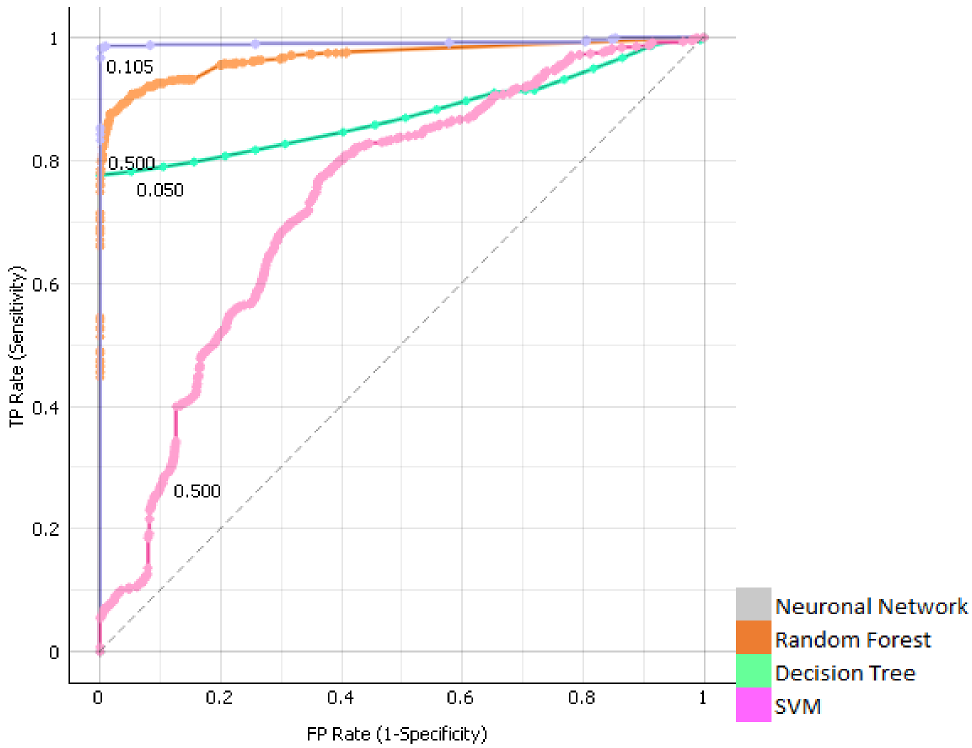

Figure 3. ROC curve

Figure 3 shows the ROC curve of the 4 Proposed ML models . _ According to the diagnosis of the learning of the models is evaluated the performance and precision , for this 2 key concepts to take are explained in bill in model choice , these are sensitivity and specificity . _ _ (Ojeda 2022)

Sensitivity; Explain the percentage of orders earrings that are recognized and learned by the ML model.
Specificity; Explain the percentage of orders cared for who are recognized and learned by the ML model.
The neural network model has the best sensitivity which generates an area under the curve Quite broad , i.e. the amount of BackOrder found in the real data it resembles the prediction of the model notoriously .

$$\text{Sensitivity} = \frac{VP}{VP + FN} = \frac{505}{505 + 11} = 97.87\%$$

$$\text{Specificity} = \frac{VN}{VN + FP} = \frac{2264}{2264 + 2} = 99.91\%$$

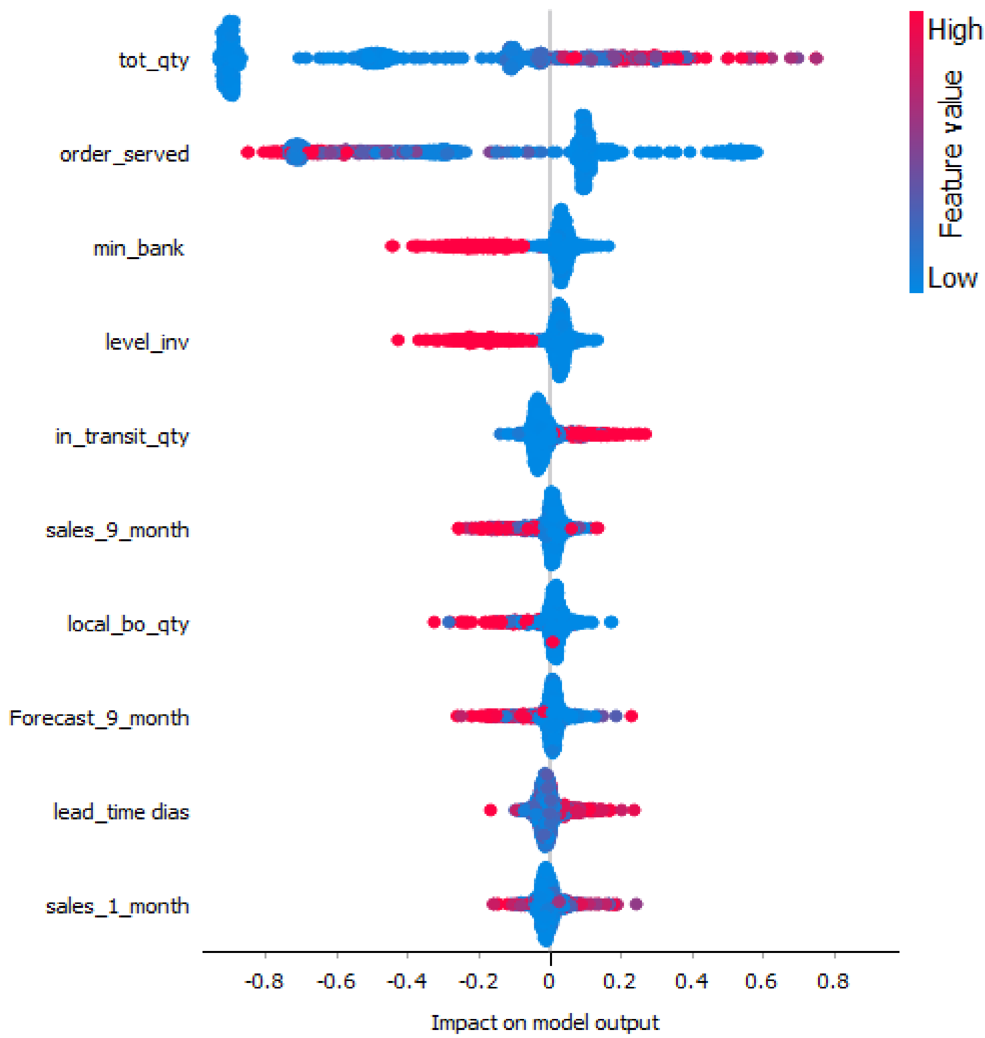

Figure 4. Impact value of the variables

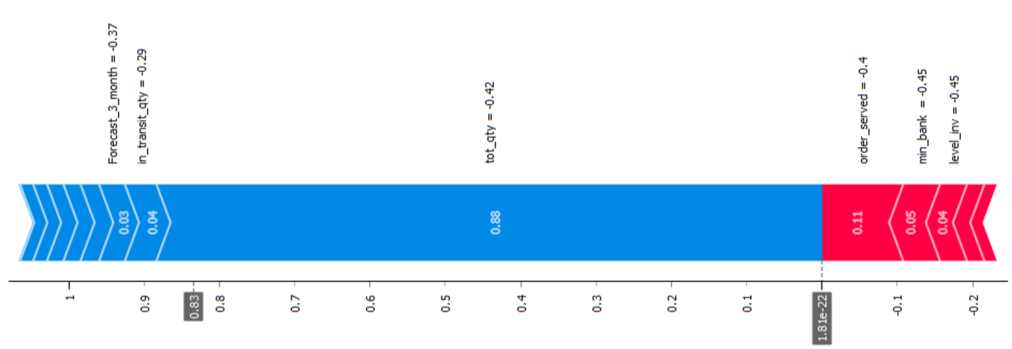

Figure 5. Impact value of the variables on the objective

5.3 Proposed Improvements

For the BackOrder analysis, four artificial intelligence models were compared, including Neural Network, Support Vector Machine (SVM), Decision Tree and Random Forest. Table 2 shows the prediction metrics according to the parameters used in each model, detailing the accuracy, the area under the curve and the confusion matrix, choosing among them the model with the highest prediction accuracy compared to the real data, the neural network reached an area under the curve of 0.994, nus obtaining the best ROC curve compared to the other models as can be seen in Figure 3, ie. the model trained adequately and was the most accurate, reaching a sensitivity of 97.87%. Also according to table 2 the SVM model reached an accuracy of 79.7%, decision tree 96.1% and random forest 96.4%, while the neural network achieved an accuracy value of 99.5%, another important metric is the recovery where the SVM model presented a value of 56.2%, a low value compared to the other models that exceeded 95%.

Figure 4 shows the variables that have a considerable impact on the increase of BackOrder, the variables colored in red are those that increase the value of pending orders, that is, those that have the greatest influence on order delays, while the variables colored in blue represent those logistic variables that have the least impact on the target variable (Back Order). Figure 5 shows in a bar chart the percentage of each variable influencing back orders.

5.4 Validation

In addition, the loess regression analysis is presented, where the relationship of the most important variables of the model is evaluated, thus creating a graph with the loess function that represents the behavior of the correlation of variables. As shown in Figure 6, there is an irregularity in the curve of the variables Order served and Total qty with a value of 85.4%-, since, for there to be no Backorders, the ideal is that the value obtained is 1. On the other hand, there is a great relationship between the minimum stock with the current inventory, this is due to the fact that the company decided to stock up with large quantities in order not to have inventory breaks in the SKUs that have continuous sales, thus seeking to prevent the accumulation of backorders, however, storage costs and inventory turnover of some codes are not taken into consideration, which is why there is a good relationship but not in an efficient manner

As a result of the interpretation of the results, it is deduced the low order attention with respect to what was ordered by the customer, due to a low inventory level and low safety stock according to the variability maintained by the 51 SKUs in the study, in addition, the model specifies the high rate of short-term sales and a very high inventory in transit. For decision making in the good supply process should be considered to evaluate the most important variables that have the greatest impact on the accumulation of pending orders, so it is necessary to evaluate the current stock in the warehouse and perform an analysis of rotation by product to determine whether the stock that is kept in the warehouse is adequate or is consistent with the demand, i, should not be ordered too much or too little, since in both situations' storage costs rise, then you must plan the correct supply based on the management of a new inventory policy.

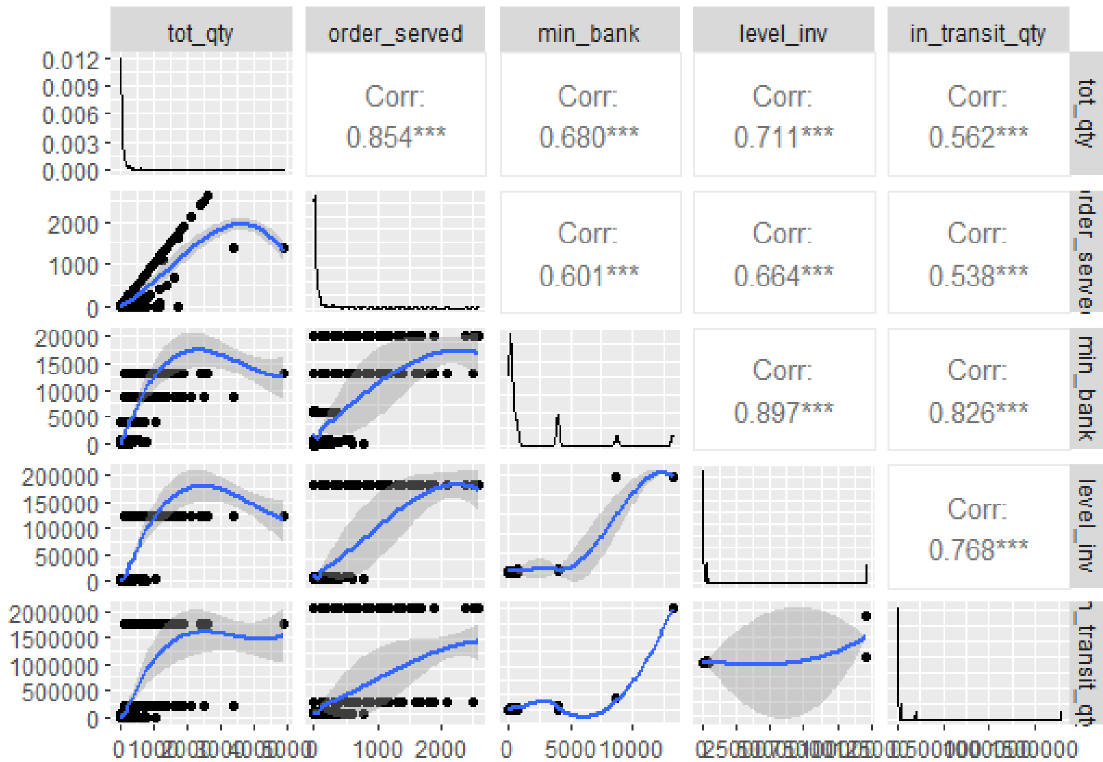

Figure 6. Loess regression

Graph 3.4 shows the correlation of the main logistics variables that influence _ in the model , through the smoothing model _ exponential .
my_fn <- function (data, mapping, method="loess", ...)
ggpairs (dataset [1:5], lower = list (continuous = my_fn))

6. Conclusions

Today's companies seek maximum efficiency in all their processes in order to satisfy the requirements of their customers and thus achieve maximum responsiveness. To this end, good inventory management is considered important for decision making and to face uncertain futures. One of the main problems that affects the level of, service is the accumulation of pending orders due to stock shortages, thus generating additional management costs and penalties that in most cases companies are obliged to pay for non-compliance. In the present work, ‘the modeling and comparison of performance indicators in the prediction using four classifiers in the Orange software was carried out, An analysis and interpretation of the main indicators is then performed to choose the best classification model. According to the training of the data by the neural network model, we obtained as results those variables that affect the decision making at the time of managing the supply process.

Thus, it is considered important to establish an inventory policy according to the sales to determine the lot to order and to propose a safety stock according to the variability of the orders. Then it can be determined that in some SKUs that have trends and seasonality in sales, the sales forecast can be calculated with the statistical model of triple additive exponential smoothing: in this case it is necessary to clean the data on some dates due to anomalous sales peaks or factors that lead to an important alteration in the orders. However, with the application of an artificial intelligence model, the data cleaning procedure was optimized because the model learns from previous events and transforms the atypical sales.

References

Adadi , A., and Mohammed B. , Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), *IEEE access* , , Vol. 6, p. 52138-52160 . 2018 https://doi.org/10.1109/ACCESS.2018.2870052

Butler, K., Davies, W., Cartwright, H., Isayev, O. and Walsh, A., Machine learning for molecular and materials science, *Nature* , vol. 559, not 7715, p. 547-555. 2018. https://doi.org/10.1038/s41586-018-0337-2
Hajek, P., and Mohammad, A., A profit function-maximizing inventory backorder prediction system using big data analytics, *IEEE Access* , , vol. 8, p. 58982-58994 . 2020. https://doi.org/10.1109/ACCESS.2020.2983118
Higa , H. , Demand Forecast Analysis for Power manage the Inventory applied to the health sector , Available: https://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/14099/Higa_Shimabukuro_Hugo.pdf?sequence=1&isAllowed=y, Jan, 2021
Islam, S., and Amin, S., Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques, *Journal of Big Data* , , Vol. 7, not 1, p. 1-22. 2020. https://doi.org/10.1186/s40537-020-00345-2
Manterola , C., Quiroz, G., Salazar, P., and García, N., Methodology of study designs most frequently used in clinical research, *Revista medical Las Condes Clinic* , vol, 30, no 1, p. 36-49, 2018.
Marroquin J., Garcia Y. and Taquia J., Machine Learning: Comparison of algorithms for determining water quality in the Rímac river. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* , , vol. 12, not 12, p. 552-572, 2021. Available: https://turcomat.org/index.php/turkbilmat/article/view/7406
Namazkhan , M., Albers, C., and Steg, L., A decision tree method for explaining household gas consumption: The role of building characteristics, socio-demographic variables, psychological factors and household behaviour . *Renewable and Sustainable Energy Reviews* , , Vol. 119, p. 109542. 2020. https://doi.org/10.1016/j.rser.2019.109542
Moshtagh , M., and Taleizadeh , A., Stochastic integrated manufacturing and remanufacturing model with shortage, rework and quality-based return rate in a closed loop supply chain. *Journal of Cleaner Production* , , vol. 141, p. 1548-1573. 2017. http://dx.doi.org/10.1016/j.jclepro.2016.09.173
Chan, SW, Tasmin, R., Aziati , AN, Rasi , RZ, Ismail, FB, and Yaw, LP, Factors influencing the effectiveness of inventory management in manufacturing SMEs. In *IOP Conference Series: Materials Science and Engineering* . Vol. 226, No. 1, p. 012024. 2017.IOP Publishing. http://dx.doi.org/10.1088/1757-899X/226/1/012024
Papernot , N., McDaniel, P., Goodfellow, I., Jha, S., Celik , ZB, and Swami, A. (2017, April). Practical black-box attacks against machine learning. *In Proceedings of the 2017 ACM on Asia conference on computer and communications security* , , pp. 506-519. 2017. https://doi.org/10.1145/3052973.3053009
Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* Vol. 1, not 5, p. 206-215 . , 2019, https://doi.org/10.1038/s42256-019-0048-x
Salgado, D., Sample probabilistic and non- probabilistic . *AUTONOMOUS UNIVERSITY OF THE STATE OF MEXICO, Mexico.* Available: https://ri. uaemex . mx/bitstream/handle/20,500 , , vol. 11799, p. 108928. 2019
Tang, Z., & Ge, Y., CNN model optimization and intelligent balance model for material demand forecast. *International Journal of System Assurance Engineering and Management* , p. 1-9, 2021,. https://doi.org/10.1007/s13198-021-01157-0
Zhang, M., Shi, L., Zhuo , X., and Liu, Y., A Bullwhip Effect Weakening Approach Based on VMD-SVM Algorithm under the Background of Intelligent Manufacturing. *processes* , vol. 9, not 11, p. 1957. 2021,https://doi.org/10.3390/pr9111957
Ojeda, M., Sensitivity and specificity of a Test for the detection of Candida albicans, Available: http://hdl.handle.net/20.500.12495/8787 , Jun, 2022
Lundberg, M., Erion , G., Chen, H., DeGrave, A., Prutkin , M., Nair, B., and Lee, I., From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* , , vol. 2, not 1, p. 56-67. 2020. https://doi:10.1038/s42256-019-0138-9
Khan, M., Qayoom , A., Nizami , M., Siddiqui, M., Wasi , S., and Raazi , S., Automated Prediction of Good Dictionary Examples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning , and Word Embedding-Based Deep Learning Techniques. *Complexity* , Vol. 2021. https://doi.org/10.1155/2021/2553199
Fonseca, B., Nogueira, H., and Ravetti , G., A hybrid Lagrangian metaheuristic for the cross-docking flow shop scheduling problem. *European Journal of Operational Research* , , Vol. 275, no 1, p. 139-154, 2019. https://doi.org/10.1016/j.ejor.2018.11.033

Acknowledgments
Thanks to the industrial engineering career for all the teachings received during our academic studies and to the professors of the faculty.

Biographies

Panduro Lope Jamil Venturo is a candidate to receive the title of industrial engineer from the Faculty of Engineering and Architecture of the University of Lima, Lima, Peru.

Pumayauri Hidalgo Sebastian Emilio is a candidate to receive the title of industrial engineer from the Faculty of Engineering and Architecture of the University of Lima, Lima, Peru.

Garcia-Lopez Yvan Jesus is PhD (c) in Engineering and Environmental Science, UNALM, "Master of Business Administration" from Maastricht School of Management, Holland, and master's in strategic business administration from Pontificia Universidad Católica del Perú. "Master of Science" in Computer Science, Aerospace Technical Center - Technological Institute of Aeronautic, Brazil. Stage in Optimization of Processes and Technologies, University of Missouri-Rolla, USA, and Chemical Engineer from the National University of Callao. Specialization Study in Digital Transformation, by Massachusetts Institute of Technology, Business Analytics, Wharton School of Management, Data Science by University of California, Berkeley, Big Data and Data Scientist by MITPro , USA Postgraduate Professor: Specialized Master from IT, MBA Centrum Católica , MBA from Calgary, Canada, and Catholic Center . Principal Consultant DSB Mobile, Executive Director of Optimiza BG, advisor to the Office of Electronic Government and Information Technology (ONGEI) - PCM, Managing Director of Tekconsulting LATAM, Executive Director of Optimiza Business Group, Ex- Vice Dean of Information Engineering of the University del Pacifico , Former Information Technology Manager of "MINERA CHINALCO PERU" Subsidiary of the Transnational Aluminum Corporation of China, Beijing, China. Former Manager of Systems and Communications of Maple Energy PLC, Director of Information Technology of Doe Run Peru SRL, Project Manager in implementation of ERP SAP, E-Business Suite - Oracle Financial and PeopleSoft. Process Analyst in transnational companies Fluor Daniel Corporation-USA, PETROBRAS- Brazil , Petroleos del Perú. He has more than 25 years of extensive experience in the management of investment projects, execution, and commissioning in Peru, Colombia, USA, Brazil, China..