Towards a Predictive Model that Supports the Achievement of More Assertive Commercial KPIs Case: Wood Trading Company

Jhon Walter Tavera Rodríguez(✉)

Grupo GIDITIC, Universidad EAFIT, Medellín, Colombia
jtavera1@eafit.edu.co

Abstract. This article presents a predictive model to determine possible causes of commercial results in a company that commercializes products and services for the furniture and wood industry in Colombia. To achieve this, a literature review was carried out to identify analysis strategies and new technologies that could influence the proposed model. Then, using the CRISP-DM methodology, the main variables and indicators that make up the business model and the problems associated with decision-making were identified, with the aim of predicting and optimizing their KPIs, improving performance metrics and achieving an increase in commercial benefits. A case study was carried out with a data set of 99,972 records collected between 2020 and 2023, which facilitated the application of variable selection techniques to identify the most influential in the prediction. The model was developed using algorithms such as decision trees, random forests, and logistic regression. Once the model was trained, it was determined that the random forest regression algorithm with the Out-of-Bag validation method and an R2 of 94.1% provided the best results and delivered the highest sales prediction. In testing, the model showed that it was influenced by variables such as average invoice value, number of invoices, available inventory, and order fulfillment. These findings expand decision-making capacity by defining which variables must be controlled to improve results. In conclusion, The machine learning-based predictive model can identify potential causes of business outcomes and improve the accuracy of decisions at a strategic level in the area of timber trading. However, it is suggested to complement it with other variables to obtain an even more precise diagnosis.

Keywords: Balance Score Card · Fuzzy logic controllers · Machine Learning · Root Cause Analysis · Artificial intelligence

1 Introduction

One of the common problems that companies must solve in their commercial process is to determine with a high degree of certainty what were the causes

M. Tabares et al. (Eds.): CCC 2023, CCIS 1924, pp. 350–366, 2024.
https://doi.org/10.1007/978-3-031-47372-2_28

that affected the results in the key business indicators and thus make strategic decisions that improve said results. In the past, diagnostic analysis was entirely manual, requiring an analyst's skills to identify anomalies, spot patterns, and determine relationships. Especially in the last five years, medium and large companies are facing new opportunities in the markets, which leads them to invest large sums of money in their transformation and improvement. However, this leads them to face aspects such as greater complexity of their business processes, the diversity and volume of information. This means that even the most experienced business analysts cannot guarantee the coherence and consistency required to achieve a good level of so-called "business intelligence" and consequently continue to use traditional decision-making processes or models based on experience. The experience-based decision-making process is influenced by the personality and mental model of each individual who participates in it. In some cases, it can be said that the causality analysis depends on the observer, cognitive abilities and their mental model are limiting in the diagnostic analysis developed by people, which directly impacts the KPI (Key Performance Indicator) thus becoming vulnerable to its result not being correctly understood to optimize resources and achieve a better return on investment (ROI). These types of risks can be mitigated with the application of analytical techniques, chosen from a series of algorithms, to determine the causality of indicator results and identify independent variables that companies can adjust to achieve positive change. This article presents a predictive model to support the achievement of more assertive KPIs - Case: Wood Trading Company presents a predictive model to determine possible causes of commercial results in a company that commercializes products and services for the furniture and wood industry in Colombia. To achieve this, a literature review was carried out to identify analysis strategies and new technologies that could influence the proposed model. The methodology used was CRISP-DM through which the main variables and indicators that make up the business model and the associated problems to make assertive decisions were identified. Likewise, the necessary data from the information systems that could lead to an accurate prediction were identified. The predictive model will make it possible to more accurately identify the causes of the results of corporate indicators of the commercial process, in a more precise and truthful way, to make better decisions and improve the results of the indicators, which will impact benefits for organizations. For this model, the main variables that make up the commercial model were identified. The scope of the work was defined with commercial data between the years 2020 and 2023, in which 9 predictor variables and one dependent variable were identified. The model was made using regression algorithms, decision trees and random forests, and finally the model was evaluated using the confusion matrix, precision, F1 Score, Recall and cross validation. The article is organized into five sections. Section 2 presents a review of the literature associated with diagnostics using Machine Learning. Section 3 presents the methodology used in this research. The development of the model, including the results obtained and the evaluation. Finally, Sect. 4 presents a discussion with the existing literature, the conclusions and some future work.

2 Literature Review

To develop a conceptual understanding of diagnostic data analysis on key performance indicators (KPIs) and the influence of modern techniques such as artificial intelligence with machine learning on the analysis process, a systematic review of the literature was conducted. The process focused on articles that used machine learning in the analysis of performance indicators. Accordingly, searches for published studies and relevant use cases were carried out using online databases. Since approximately 2017, different investigations have been published around how KPIs are influenced by Machine Learning techniques, some of them are:

[5] present a case study to measure the productivity of manufacturing wiring harnesses within the automotive industry, using machine learning algorithms in KPI prediction. The data was provided by the cutting department of an automotive wiring company. In this case study, the dataset consists of 7 variables, 1,917 observations, and 1 objective (OEE value, Overall Equipment Effectiveness), a metric to identify losses, compare progress, and improve the productivity of manufacturing teams (i.e., eliminate waste). The OEE calculation gives managers the possibility to monitor their process and identify the main losses that reduce the effectiveness of the machine. The objective of this article is to develop a model based on machine learning, which is based on historical measurements of the process, capable of predicting the estimated value of OEE. The predicted OEE value will allow managers the ability to assess the effectiveness of the equipment and therefore examine the inputs to the production process to identify and eliminate relative losses. This will form a decision support tool designed for the purpose of helping managers prevent different types of losses and then act and react according to the given situation to improve and maximize production effectiveness. In the case study, they first tested different linear regression models, namely multiple linear regression, polynomial regression, support vector regression (SVR) and decision tree regression optimized support vector regression. Second, set learning techniques are manipulated with tree-based models. Random forest is tested for the data set and XG-Boost for the boost set. Finally, a deep learning predictor is built to generate a robust model with a high level of performance. The result shows that the use of the cross-validation technique with a relevant division of the data provides more reliable, accurate and unbiased predictors. This comparison was made based on the MAE metrics, MAPE and RMSE and confirmed by statistical techniques such as ANOVA and standard deviation tests. Regarding the random forest, XG-Boost is implemented under two configurations, one that uses simple data division (XGB) and another that uses cross validation (XGBCV), this model was accurate, more efficient and reliable than the others. It is worth mentioning that the result of the investigation mentions that the OEE is just a case study that could be generalized to other KPIs, and the experiment could also be carried out in any other industry, ensuring the necessary changes and adaptations.

[8] use statistical and machine learning methods in an experiment to detect anomalies that may occur in internet service servers in a data center. The goal of the project is to ensure the stability of web services, by monitoring various

key performance indicators (KPIs) to judge whether the web service is stable or not. This article divides the problem into two parts. First, time series analysis methods such as Triple Order Exponential Smoothing (Holt-Winters) and the ARIMA model and regression-based machine learning techniques such as Gradient Boosting Regression Trees (GBRT) and Long Short-Term Memory (LSTM) to predict the value at the next point in the time series. After that, they set the anomaly detection rule. Finally, the predicted value is compared with the actual value to determine if the current point is outlier or not. KPI data corresponds to timestamp, value that was saved at 1 min intervals, means that each hour had 60 data points and each day had 1440 data points. They chose a KPI representative of a search engine that shows search page view (PV). This is one of the most widely used indicators for evaluating website traffic, monitoring the trend of change in PV and analyzing the reasons for its change. In the experiment output they used MSE to evaluate a binary classification and precision, recall and F1 score to evaluate the models. In the project report it is concluded that the models perform well in detecting anomalies, especially Gradient Boosting Regression Trees (GBRT), have the highest recovery and the shortest training time. Machine learning is an efficient and promising method for anomaly detection for KPIs. In the conclusions of the experiment they propose to test other methods of machine learning, especially deep neural networks.

[10] explores how to apply machine learning in two different domains; in predictions of defect inflow and accumulation of single product defects at Ericsson AB company and the state level of parameters used in automotive projects at Volvo Car Corporation (VCC). At VCC, the propulsion unit software department is being investigated. The KPI that relates to the parameter calibration status of a project is inspected. A desired result of applying machine learning would involve a prediction about what the average state will look like in the future. These predictions can help the decision maker in the planning process. A special desire in predictions is to find underlying patterns that make the predictions accurate and reliable. Ericsson AB has established a way of incorporating its KPIs into the daily use of the organization. As there are many KPIs to investigate, one stakeholder suggested the defect inflow rate. This is then expanded by also investigating the Defect Accumulation KPI. The linear regression method is applied to different aspects of the Health KPI in VCC. The rolling time frame is applied to the KPIs investigated in the remaining investigation cycles. This thesis investigated a KPI VCC form regarding the status of calibration variables used in hardware components. The first approach was to use linear regression on the data. The second approach was to use a rolling time frame to predict new values given a certain lag. The second stage of this project was the KPIs investigated at Ericsson AB. The two KPIs were Defect Inflow and Defect Accumulation. Here, the approach used in VCC was repeated using the moving time frame. The results of these predictions showed that the best approach to predict a week ahead was to use the previous value as the prediction, as this resulted in the minimum error of 1%. The best result was achieved using a rolling window and the instance-based learning KNN on the defect input stream. The Defect

Entry KPI at Ericsson AB was the KPI that proved to be the best and was improved with predictions.

[9] present a study using KPIs from the perspective of Balance Score Card (BSC) customers to estimate the success of a new web product using machine learning. Specifically, the estimation was achieved by applying an Artificial Neural Networks (ANNs) algorithm using the Python programming language. In more detail, this paper examines the success of an online product (financial services platform) using KPIs and an ANN model to determine if a potential customer is likely to use the platform. Data collection was carried out using a questionnaire. The sample was obtained by snowball sampling (sharing it on social networks), in an effort to be as large and representative as possible. The questionnaire that was created consisted of two (2) groups of questions: 1) demographic questions and 2) questions related to the scope of the platform's services. The purpose of this article is to demonstrate how a start-up company can estimate the KPI responsible for the success of an online platform before it is launched on the market. The estimation is done using AI, using the KPIs derived from the perspective of BSC customers. For the purposes of this study, the potential development of an online platform by a technology company that provides financial services is explored. This research addresses the contribution of Machine Learning in the design of a strategy for optimal decision making. This occurs as a result of the ANN's ability for pattern recognition and its strength to mimic the neural networks of the human brain. This procedure is currently much more efficient thanks to faster computer processing. From the analysis of the data using the ANN, the following results were obtained. Of 122 participants, 48 chose option number five (5), making it the most popular, indicating that they would use a financial services platform to the extent of 5: A lot (with a range of 1: Not at all - 5: A lot). The result of the questionnaires revealed that 7/10 people would use this online platform, indicating that it would be successful. However, using Machine Learning, it was found that with an accuracy of 91.89%, the company's product will be successful. In addition, given that the threshold that would determine whether or not the new product would be built was set at 60%, and the results revealed that 70% of people would use this online platform, it is concluded that the technology firm in this study will continue with new product development. It is also worth noting that the presence of high precision in the Machine Learning model indicates that the development of the new product will have less financial risk compared to not using Machine Learning for this analysis. In conclusion and based on all the findings of this study, the application of Machine Learning in business is very important. For optimal results, Companies should strive to combine accurate interpretation of the results by qualified analysts and proper data management techniques to be applied before building the models, which can help companies improve their production and economic growth. Additionally, automation frees up more time for analysts to focus on other aspects of the business plan, enabling them to make better decisions. It also reduces costs and improves the customer experience.

[7] present a time series anomaly detection method based on correlation analysis and hidden Markov model (HMM) in the field of intelligent operation and maintenance. You can identify abnormal KPIs within a set period from a large number of KPIs in a system and the transfer status between them. Correlation analysis is used to obtain the correlation between abnormal KPIs in the system, which reduces the false alarm rate of anomaly detection. In the field of smart operation and maintenance, KPI anomaly detection for a multi-index system is difficult. In system KPIs, there can be a temporal correlation between two indicators; that is, when one KPI is abnormal, it will cause similar fluctuations in the trends of other KPIs in a short time. Due to the presence of numerous trailing indicators and complex structures in the distributed system, the fluctuations will continue to spread to more KPIs. This causes more KPI anomalies throughout the system, making anomaly detection and root cause analysis very difficult. Without using correlation analysis, it is very difficult for OyM personnel to analyze and infer the correlation between a large number of complex KPI anomalies. In the anomaly detection method discussed in this paper, a 1D-CNN-TCN prediction model is first proposed to predict the KPIs and obtain the residual sequence to detect the possible abnormal KPIs. This model combines the local variable acquisition capability of convolutional neural network (CNN) with the temporally dependent variable acquisition capability of temporal convolutional network (TCN) to improve prediction accuracy. The residual sequence of abnormal KPIs can highlight the abnormal segment in each KPI, so that the correlation analysis is not affected by the original fluctuation of the KPIs and thus the accuracy of the correlation analysis is improved. The experimental results show that the F1 score of the correlation analysis method in this paper is also the best. The HMM parameters are confirmed based on the correlation matrix. After training the HMM, other KPIs that can cause a KPI abnormality are found, reducing the time required for OyM personnel to find a large number of abnormal KPIs. According to the results of the F1 score evaluation, the optimal F1 scores of abnormality detection and time order are 0.94 and 0.90, respectively, indicating that this method has a good effect on abnormality detection. The method in this document can further determine the influence relationship between KPIs by getting abnormal KPIs, which can help to build a fault propagation diagram and help OyM personnel to perform quick troubleshooting. However, the method of this paper still has some limitations, and the specific threshold of the correlation analysis method still needs to be adjusted according to different environments.

[4] investigate an approach to detect anomalies in financial data, using behavior change indicators (BCIs). The application of the BCIs proposed in this study can be described as a KPI time series analysis. From a data science standpoint, tracking financial KPIs is critical to managing activity and changes in KPIs over time. Performance management assessment requires an understanding of the characteristics of changes in KPIs over time. Based on this assumption, three types of BCI have been defined to detect and assess changes in traditional KPIs in time series: absolute change indicators (BCI-A), relative change indicators

(BCI-RE proportion indicators) and delta forex indicators. (D-BCI). Financial KPI thresholds (normative values) help to assess the state of a company's operations during the financial period, allowing the creation of templates (patterns) that the software system uses to perform such an assessment (for example, zone safe zone, risk zone, disaster zone). Another problem is the financial limits (normative values, thresholds) of the BCIs, which must be set separately for each KPI. By setting BCI thresholds, templates can be created. The software system would automatically assess BCI values and the risks of changes in the KPI time series and would detect KPI changes that indicate anomalies in financial data. Examples of raw accounting data (attributes) used in the experiment include a number of variables including: DBID, FinancialYear, Total Equity, Total Liabilities, Total Assets, Non-current Passive, Current Passive, Revenue, OPEX, Other Income, Financial Income, Taxes, Gross Profit, Operating Profit, EBITDA, Net Profit, Gross Margin%, EBITDA%, NetMargin%, ROA YTD, ROE YTD, Cash Position, Cash Ratio, Working Capital, Current Ratio, Acid Test, Debt/Equity Ratio, Debt/Asset Ratio, OPEX%, Intangible Fixed Asset Ratio, AR, AP and inventory. The advantages of using BCI in financial data analysis can be summarized as follows: BCIs provide a quantitative estimate of KPI changes over a defined period. BCIs highlight KPI changes that are almost impossible for a person to see and understand when analyzing KPI values or just their changes. If the BCI exceeds a certain threshold, such a change in the value of a KPI is one of the most important signs of a change in a company's business decisions or processes or human error. It is almost impossible for an individual expert to detect and understand anomalies simply by analyzing the wealth of financial KPI data. Using BCIs to track changes to KPIs helps users see suspicious trends in some KPI changes (indicating potentially suspicious data) and reduces the amount of data that needs to be analyzed.

[2], developed a method to determine and calculate the performance of employees in virtual environments and access the best talent as a form of competitive advantage using machine learning through the use of methodologies such as the Scikit-learn Multi-Output-Regressor function. Due to certain changes in the formulation of virtual leadership styles, such as the trend towards physically dispersed work groups, new research on the role and nature of team leadership in virtual environments has become necessary more challenging than managing people on the site. The total staff of the company considered is currently 51 employees made up of 78% men and 22% women. The department is made up of vertical teams. Each has a team leader and a team leader responsible for giving directives, setting priorities and goals for the team. Several inducer algorithms were discussed in this project, such as linear regression or a support vector regression (SVR) model with a linear kernel models data in a completely different way than an instance-based learner like the k-neighbor algorithm closest (KNN). Scikit-learn implements a function called GridSearchCV, which is short for grid search cross-validation, which can be used to determine the optimal algorithm along with a number of specific hyperparameters. GridSearchCV sets up a grid of each algorithm with all possible combinations of the given parameters

and their respective values, and using cross-validation and R2 scoring, finds the optimal solution.

[6], used machine learning to predict the performance of key indicators for construction projects. The different aspects of the projects are evaluated by key performance indicators (KPIs) that are used to monitor and control construction projects. In the project, a novel framework was proposed to qualitatively measure and predict six important construction project KPIs using the neuro-fuzzy technique. To map the KPIs of three critical stages of the project to the KPIs of the entire project, neurofuzzy models were developed. In the proposed framework, the neuro-fuzzy technique was applied to forecast the KPIs of the entire project automatically from the data. The neuro-fuzzy technique is a combination of artificial neural network (ANN) and fuzzy logic, which is used to solve different research problems in construction management. In this research, the neural network (ANN) model was used to predict the KPIs of the entire project. The input of the prediction models was 18 KPIs, six KPIs for each of the three stages. The results were six KPIs for the entire project. The ANN models were developed, trained, and tested in MATLAB 2016a. Models were developed using three available training algorithms for neural networks: Levenberg-Marquardt (LM), Bayesian regularization (BR), and scaled conjugate gradient (SCG). The performance of the model was evaluated based on the coefficient of determination (R2), the mean absolute error (MAE), the relative absolute error (RAE), the root relative squared error (RRSE), and the mean absolute% error (MAPE) and each amount of error rate. Six KPIs were chosen for the frequency of their use in the literature; There is cost, time, quality, safety, customer satisfaction, and project team satisfaction. KPIs from three critical project stages (early stage, mid-stage, and late stage) were used to predict project-wide KPIs using two main techniques: artificial neural networks (ANNs) and neuro-fuzzy. In the ANN, the best model was selected by changing the number of neurons in the hidden layer. Neurofuzzy models were developed in two steps; First, the initial FIS models were developed using both subtractive clustering and FCM. In subtractive clustering, the radius of the clustering was optimized to achieve optimal precision without overfitting. Second, optimization of the initial FIS model was performed using ANN. In the neuro-fuzzy technique, 18 different models were developed, six models for each of the three critical stages of the project. The results indicate that the neurofuzzy technique using subtractive clustering performs better and has lower error values compared to the other two methods. Therefore, to predict construction project KPIs, the neuro-fuzzy technique using subtractive clustering is recommended. The proposed framework is designed to be flexible and can be applied to other countries and other types of projects.

[3], elaborates an investigation to establish the bases of a system for the automatic diagnosis of failures in mobile networks using automatic learning. With the knowledge available about the network performance data, the most common problems, and the manual troubleshooting process, the next step was to design an artificial intelligence algorithm that would perform the self-diagnosis. In this thesis, the FLC (Lotfi A. Zadeh. Fuzzy sets. Information and control)

are used as the best option, since they are easily understandable both by human experts (since they use a language close to spoken) and by machines. Fuzzy logic controllers facilitate the automatic generation of diagnostic rules and their integration with existing rules. This solution is also attractive to network operators, as its clarity makes it less confusing than other alternatives, such as Bayesian networks or neural networks. At work, whenever an expert diagnoses a mobile network problem, the expert can report it with minimal effort and store it in the system. The core of this system is a diagnostic algorithm (a Fuzzy Logic Controllers - FLC) that evolves and improves by learning from each new example, until it reaches the point where experts can count on its accuracy for problems more common. Every time a new problem arises, it will be added to the system's database, thus further increasing its power. The goal is to free experts from repetitive tasks, so they can spend their time on challenges that are more rewarding to solve. Therefore, the first objective of this thesis is the collection of a database of real failure cases. For it, A user interface for data collection is designed taking into account ease of use as a priority requirement. Once the collected data is available, it will be analyzed to better understand its properties and obtain the necessary information for the design of data analytics algorithms. Another objective of this thesis is the creation of a mobile network failure model, finding the relationships between network performance and the occurrence of problems. The acquisition of knowledge is done through the application of analytical algorithms on the collected data. A KDD process is designed that extracts the parameters of a fuzzy logic controller and is applied to the collected data base. Finally, this thesis also aims to carry out an analysis of the Big Data aspects of the Self-healing functions, and take them into account when designing the algorithms. The model of network failures obtained in this work constitutes a valuable contribution that helps to understand the relationships between performance measures and modeled failures. Furthermore, the model can be used to generate highly realistic emulated data. The procedure used to extract the model can also be used to model new problems and in different networks. Another contribution of this work is the study of the application of FLCs to diagnosis and the design and implementation of data mining algorithms that adjust a fuzzy logic controller based on a database of diagnostic cases. The extracted fuzzy logic controller (FLC) achieves a high success rate in the diagnosis of the network under test; and the method used to adjust it can be used in other networks. In addition to being useful for diagnosis, the rules are a good source of information about the nature of the observed problems, since they are given in a language close to human.

[11] propose a generic causality learning approach for tracking and monitoring production cycle time. Therefore, a causality analysis of the KPI values is presented, as well as a prioritization of their influencing factors to provide decision support. The purpose of this article is to provide an approach to make predictions and diagnoses in order to detect abnormal situations and identify their causes. For this, Bayesian Networks (BN) are used, which aim to identify the factors that affect the KPI addressed. ANNs are used to solve complex problems

and enable learning and modeling of complex, non-linear relationships between inputs and outputs. To implement the proposed data analysis methodology, the authors have constructed a representative summary data set to validate that the experiments are correct. This dataset respects, in a very flexible way, a certain amount of causality rules that have been previously defined to be compared with the resulting causality links. The use case addresses one KPI: production cycle time. The rest of the data set is made up of variables that may or may not affect the addressed KPI: the day of the week, the time zone, the month, the indoor temperature, the operator's heart rate, their stress level, the default number and the level of training. The objective is to identify, among this data set, the variables that affect the production cycle time and prioritize them. In order to build a realistic and more relevant use case, the authors have ensured that the dataset does not exclusively follow these rules. Basically, this KPI is calculated from two pieces of information: the remaining production time and the number of units produced. Given this, if a deviation is detected, both pieces of information do not give answers either to understand how this happened or to trace the root causes. Use a constraint-based algorithm (Peter and Clarck (PC) algorithm) to define the structure of the BN that will allow causal links to be identified. The PC algorithm is a constraint-based algorithm that starts with a complete undirected graph and removes edges between pairs that are not statistically significantly related by performing conditional independence (CI) tests. The proposal is currently in its early stages of development and many aspects need to be addressed, such as testing and benchmarking of other algorithms and BN structure learning tools.

3 Methodology

The first step in this research methodology was to define the objectives and goals of the project. Subsequently, the CRISP-DM (Cross Industry Standard Process for Data Mining) case study methodology [1] was carried out, which allows us to establish a framework structured to measure and diagnose the causes of the results in business indicators. This gives us support in decision making by identifying relevant actions to improve the value of the KPI. The process includes the identification of the main KPIs of the project through the review of the literature, the knowledge of the business and the opinion of experts, as well as the application of a set of machine learning techniques. The methodology is structured in a hierarchical process, composed of tasks described at different levels of abstraction, ranging from the general to the specific. The process is divided into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Implementation. The activities carried out in each of these phases are described below.

3.1 Business Understanding

The investigation was carried out in a company dedicated to the commercialization of products and services for the furniture and wood industry. The company

currently has a nationwide presence and has different operation centers that handle customer orders and services. In addition, it has various lines of business that range from the marketing of raw products such as boards and finished products such as doors, floors, furniture and design services. Currently, the company calculates the results indicators taking into account variables such as quantity delivery compliance, number of orders delivered on time, delivery time in days, number of customers, quantity sales value, number of invoices and number of orders. A group of business experts periodically analyze the causes of their results and have identified that it is not possible to guarantee an accurate diagnosis and accurate decision-making that reflects the correct operation of the company. For this reason, they consider that it is necessary to involve new variables in decision making, such as the average value per invoice and amount of inventory. These new variables are expected to help identify in advance what is causing current misjudgment. Other variables that could be important to consider are: availability of personnel for the process, square meters of the operation centers, location of the operation centers, service level qualification, and marketing and advertising, however, they are in process data collection by the company.

Each of these variables is defined below.

- Compliance delivery in quantities: This characteristic measures a variable that can impact sales and corresponds to the level of compliance in the quantities delivered requested by the customer.
- Orders delivered on time: Corresponds to the number of orders delivered on time, a variable that comes from the company's business intelligence system.
- Delivery time in days: Variable that corresponds to the delivery time of the orders in days.
- Number of customers: Independent variable that corresponds to the number of customers that exist per operation center.
- Quantity sales value: Corresponds to the sum of the quantities sold.
- Number of Invoices: Independent variable that corresponds to the number of invoices issued in the period.
- number of orders: Independent variable that corresponds to the number of orders generated in the period.
- average value per invoice. A variable of the average invoice value was created, at the granularity level of year, month, operation center and product line; This is because the average value is a feature that helps us in the analysis of sales behavior.
- Inventory Quantity: Independent variable that contains the quantities of inventory available at the end of the period (Figs. 1 and 2).

Author	application sector	Predictor Variables	Objetive	Techniques Used
Mazgualdi, Masrour, Hassani, & Khdoudi, 2021	Manufacturing in the automotive industry	Number of orders, days of production, number of losses.	Overall Equipment Effectiveness (OEE)	Random Forest - XGBCV-Boost
Shi, He, & Liu, 2018	Internet Service Companies	Timestamp, website traffic	Internet service stability	Gradient Boosting Regression Trees (GBRT)
THORSTROM, 2017	Manufacture of electronic devices	parameter calibration status, defect input flow rate	Defect flow prediction	KNN learning
Tagkouta, Psycharis, Psarras, Anagnostopoulos, & Salmon, 2023	Companies with online sales	scope of platform services, demographic data.	Estimate the success of a new web product	Artificial Neural Networks (ANNs)
Zijing, and others, 2021	Smart Maintenance	failure propagation, residual sequence of abnormal KPIs	Identify abnormal KPIs	Convolutional Neural Network (CNN)
Lopata, and others, 2022	Financial Process	DBID, FinancialYear, Passive, Revenue, OPEX, Other income, Financial income, Taxes, Gross profit, Operating profit, EBITDA, Net profit, Gross margin%, EBITDA%, NetMargin%, ROA_YTD, ROE_YTD, Cash_Position, Cash_Ratio, Working_Capital, Current_Ratio , Acid_Test, Debt/Equity Ratio, OPEX%.	Detect anomalies in financial data using behavior change indicators (BCI).	Time series analysis,
Paxleal, 2020	Work companies in virtual environments	online leadership styles	employee performance	Scikit-learn Multi-Output-Regressor.
Fanaei, Moselhi, Alkass, & Zangenehmadar, 2018	construction companies	cost, time, quality, safety, customer satisfaction, and project team satisfaction.	Project performance	Artificial Neural Network (ANN) and Fuzzy Logic
Khatib, 2017	Mobile networks	failure occurrence	network performance	Fuzzy Logic Controllers (FLC)
AMZIL, YAHIA, KLEMENT, & ROUCOULES, 2021	production companies	Production day, time zone, month, indoor temperature, operator heart rate, stress level, training level.	production cycle time	Bayesian networks

Fig. 1. Summary of Findings

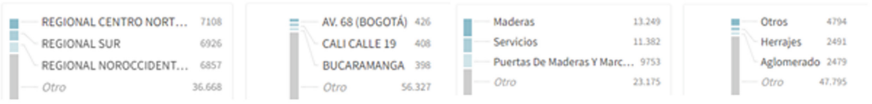

Fig. 2. Independent Categorical Variables

3.2 Data Understanding

To achieve the objective set out in the company's need, data was extracted from information systems such as sales, logistics and financial. Data was taken between the month of April 2020 and the month of March 2023, since these met quality elements such as completeness, integrity and availability, required for their analysis. In addition, relevant categorical data was identified in the following data models: commercial, supply, and financial for 13 regions distributed nationwide, 230 operation centers, 14 product groups, and 73 product lines. Based on this information, it was identified that the dependent variable is the value of sales and the variables considered influential correspond to sales in units, compliance with the amounts requested by the client, on-time delivery compliance, order delivery time, number of issued invoices and average value per invoice, etc. These variables were subjected to a cleaning process and their corresponding normalization at the year, month, operation center and product line level. Finally, it was possible to have a data set with 99,972 records and 10 variables, for a total of 999,720 data.

Descriptive data analysis

3.3 Data Preparation

In this phase, new variables are created from those provided by the business. Each of the variables and their calculations are described below average value per invoice. A variable of the average invoice value was created, at the granularity level of year, month, operation center and product line; This is because the average value per invoice is a feature that helps us in the analysis of sales behavior. Compliance delivery in quantities. Corresponds to the percentage of the quantities delivered versus those requested by the client. Order fulfillment: A variable was created that stores the fulfillment of orders, comparing the number of orders delivered on time versus the orders generated (Fig. 3).

A heat map is used to show the correlation matrix between the first group of variables selected for the model, as shown in Fig. 4. The higher the positive

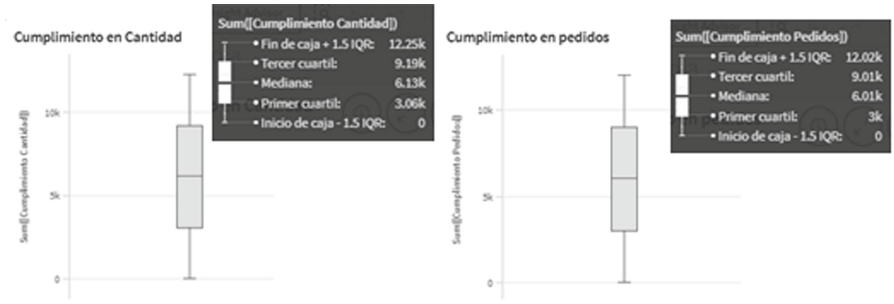

Fig. 3. Independent Variables

Fig. 4. Correlation Matrix (Color figure online)

Algorithms	variables
▪ Linear regression	▪ Average value of invoices
▪ Random forest regression	▪ Inventory Quantity
▪ XGBoost regression	▪ Number of Clients
▪ LightGBM regression	▪ Delivery time in days
▪ CatBoost regression	▪ Number of Invoices
	▪ Orders delivered on time
	▪ number of orders
	▪ Order Fulfillment

Fig. 5. Algorithms and variables selected for each technique

	Arriba	Algoritmo	R2	RMSE	MSE	MAE
☑	🏆	Regresión de bosque aleatorio	0.941	19.047.197,266	362.795.723.702.579.250	4,080,836.136
☐		Regresión CatBoost	0.847	30.641.947,001	938.928.916.008.707.000	4.261.661,994
☐		Regresión LightGBM	0.743	39.774.551.380	1.582.014.937.463.750.500	4,984,066.516
☐		Regresión XGBoost	0.901	24.686.214.478	609.409.185.271.831.600	3,966,363.104
☐		Regresión lineal	0.304	65.406.412.538	4.277.998.801.093.056.000	13.514.461,502

Fig. 6. ROC-AOC of the different models studied

correlation between the variables, the deeper the red color, for conversely, blue signifies a negative correlation.

In the figure it can be seen, in this first group of variables, that quantity delivery compliance has a high correlation with order delivery compliance. Therefore, we only take into account the order delivery fulfillment variable in the model and discard the quantity delivery fulfillment variable (Fig. 5).

3.4 Modeling

Based on the independent variables identified in the previous section, it has been proposed to use a regression model using the random forest regression

algorithm. It has been automatically selected as the best alternative based on the Machine Learning component that is part of the Qlik Cloud platform used, after evaluating different algorithms such as CatBoost Regression, LightGBM Regression, XGBoost Regression and Linear Regression.

100% of the data was used to train these models, which represent the inputs. These records were divided into two sets, one representing the training data set with 80% of the records (79,977 records), and the other representing the test data set with 20% of the records (19,995 records).

3.5 Results

Depending on the selected objective, the following algorithms were used in the analysis.

When training the model with the 10 established variables, the best algorithm was the random regression of the forest, with an accuracy of 94.1%.

Figure 6 shows the results of the different algorithms used where the random forest regression stands out with an R2 of 0.941. Machine learning prediction performances are evaluated in terms of Root Mean Square Error (RMSE) as a validation function. The results shown in Fig. 6 indicate that the RMSE ranges from 19,047 to 65,406, which shows a good response to the ML models. As expected, linear regression performed least well because its limitations lie in matching data that is not linear and predicting data that is not within the range of the training sets. CatBoost works best because it reduces any form of overtuning and classifies each attribute precisely. However, it's worth noting that Random Forest Regression outperforms its counterparts. You can appreciate the importance of the variables and their influence on the average ticket variable (Average invoice value), the number of invoices and the quantities sold, as well as the available inventory as shown in Fig. 7.

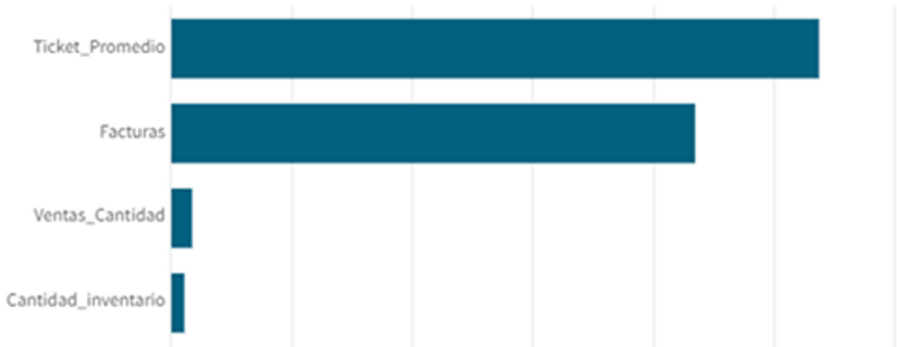

Fig. 7. Importance of the variables - random forest regression method

4 Discussion, Conclusions and Future Work

This article proposes a model to diagnose the causes of the sales results of a company in Colombia dedicated to the commercialization of products and services for the furniture and wood industry. Many variables, both internal and external, intervene in the identification of these causes, which makes it difficult to carry out a detailed analysis of the results. Diagnostic analysis based on people's experience does not guarantee accurate results and may result in personal criteria influenced by the way they view and contextualize the data, as well as by their knowledge, experience, and judgment of the circumstances surrounding the problem. Our model has shown good performance in detecting the causes of sales results, especially using the random forest regression algorithm, with an accuracy of 94.1% calculated with the cross-validation method. This means that current sales are centered around the average invoice value, which is important information for planning strategies to increase the average invoice value. These strategies can include promotions and cross-selling to improve results. A differentiating factor of this work consisted in the application of theories and methodologies that help to identify the variables that affect sales. The variables that contribute to the efficiency in a commercial process were analyzed using the pentagon model and the value chain in retail companies. This provides the necessary knowledge in the commercial function and allows an efficient response to the demands and challenges of the customer needs satisfaction cycle, thus improving commercial results in this type of business. Therefore, our proposal differs from the models mentioned in the state of the art by grouping variables that are determinant in the results obtained. However, there is still room for improvement in our research by incorporating other exogenous variables into the model. For example, the behavior of the country's construction indicators can be considered, which can generate demand for services and products. In addition, variables such as purchasing power and the effects of market strategies allow the model to be improved.

References

1. Chapman, P., et al.: Wirth: CRISP-DM 1.0: step-by-step data mining guide (2000). https://api.semanticscholar.org/CorpusID:59777418
2. Simi Paxleal, J.: Measuring KPIs of virtual teams in global organization using machine learning. Int. J. Adv. Res. Sci. Commun. Technol. 21–27 (2020). https://doi.org/10.48175/ijarsct-243
3. Khatib, E.J.: Data analytics and knowledge discovery for root cause analysis in LTE self-organizing networks. Ph.D. thesis, Universidad de Málaga (2017)
4. Lopata, A., et al.: Financial data anomaly discovery using behavioral change indicators. Electronics **11**(10), 1598 (2022). https://doi.org/10.3390/electronics11101598
5. Mazgualdi, C.E., Masrour, T., Hassani, I.E., Khdoudi, A.: Machine learning for KPIs prediction: a case study of the overall equipment effectiveness within the automotive industry. Soft. Comput. **25**(4), 2891–2909 (2020). https://doi.org/10.1007/s00500-020-05348-y

6. Fanaei, S.S., Moselhi, O., Alkass, S.T., Zangenehmadar, Z.: Application of machine learning in predicting key performance indicators for construction projects. Ph.D. thesis, Univ., Montreal, QC, Canada (2018)
7. Shang, Z., Zhang, Y., Zhang, X., Zhao, Y., Cao, Z., Wang, X.: Time series anomaly detection for KPIs based on correlation analysis and HMM. Appl. Sci. **11**(23), 11353 (2021). https://doi.org/10.3390/app112311353
8. Shi, J., He, G., Liu, X.: Anomaly detection for key performance indicators through machine learning. In: 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC). IEEE (2018). https://doi.org/10.1109/icnidc.2018.8525714
9. Tagkouta, E., Psycharis, P.N., Psarras, A., Anagnostopoulos, T., Salmon, I.: Predicting success for web product through key performance indicators based on balanced scorecard with the use of machine learning. WSEAS Trans. Bus. Econ. **20**, 646–656 (2023). https://doi.org/10.37394/23207.2023.20.59
10. Thorstrom, M.: Applying machine learning to key performance indicators (2017). https://odr.chalmers.se/server/api/core/bitstreams/bee2170c-60ec-4c2e-b3b0-f63ebf1ece00/content
11. Zapata, S.M., Klement, N., Silva, C., Gibaru, O., Lafou, M.: Collective intelligence application in a kitting picking zone of the automotive industry. In: Gerbino, S., Lanzotti, A., Martorelli, M., Mirálbes Buil, R., Rizzi, C., Roucoules, L. (eds.) JCM 2022. LNCS, pp. 410–420. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-15928-2_36