Production
ISSN: 0103-6513
ISSN: 1980-5411
Associação Brasileira de Engenharia de Produção

Arboleda-Florez, Mariana; Castro-Zuluaga, Carlos
Interpreting direct sales' demand forecasts using SHAP values
Production, vol. 33, e20220035, 2023
Associação Brasileira de Engenharia de Produção

DOI: https://doi.org/10.1590/0103-6513.20220035

Available in: https://www.redalyc.org/articulo.oa?id=396773998002

ISSN 1980-5411 (On-line version)
Production, 33, e20220035, 2023
DOI: 10.1590/0103-6513.20220035

Thematic Section - Resilient and innovative operations management

Interpreting direct sales' demand forecasts using SHAP values

Mariana Arboleda-Florez[a]*, Carlos Castro-Zuluaga[a]

[a]Universidad EAFIT, Medellin, Colombia

*marbol12@eafit.edu.co

Abstract

Paper aims: Several concerns regarding the lack of interpretability of machine learning models obstruct the implementation of machine learning projects as part of the demand forecasting process. This paper presents a methodology to support the introduction of machine learning into the forecasting process of a traditional direct sales company by providing explanations for the otherwise obscure results. We also suggest incorporating human knowledge inside the machine learning pipeline as an essential part of capturing the business logic and integrating machine learning into the existing processes.

Originality: Using explainable machine learning methods on real-life company data demonstrates that machine learning techniques are functional beyond the academy and can be introduced to everyday companies' production.

Research method: The project used real-world data from a company and followed a traditional machine learning pipeline to collect, preprocess, select and train a machine learning model, to conclude with the explanation of the model results through the implementation of SHAP

Main findings: The results provided insights regarding the contribution of the features to the forecast. We analyzed individual predictions to understand the behavior of different variables, proving helpful when interpreting complex machine learning models.

Implications for theory and practice: This study contributes to a discussion about adopting new technology and implementing machine learning models for demand forecasting. The methodology presented in this paper can be used to implement similar projects on interested companies.

Keywords

Explainable Artificial Intelligence. Machine learning. Sales forecasting.

How to cite this article: Arboleda-Florez, M., & Castro-Zuluaga, C. (2023). Interpreting direct sales' demand forecasts using SHAP values. *Production*, *33*, e20220035. https://doi.org/10.1590/0103-6513.20220035

Received: Mar. 4, 2022; **Accepted:** Nov. 29, 2022.

1. Introduction

Demand forecasting is one of the critical activities of supply chain management (Crum & Palmatier, 2003). As it refers to predicting future sales, demand forecasts support managerial decisions and operational planning throughout the supply chain (Bandeira et al., 2020). For instance, Sales and Operations Planning (S&OP) recognizes demand forecasts as an essential input for the process (Seeling et al., 2019). Demand forecasts lead the supply and operations plan and feed the business plan strategically. Figure 1 illustrates the five steps of the S&OP process and how the different functional areas of the organization are involved in every step.

During the first two steps, the Sales & Marketing, Finance, and Forecasting departments discuss and gather information about new product releases and past sales and propose a consensus demand forecast for the upcoming period. Usually, this forecasting process is seen as "[...] a combination of an extrapolation of what has been observed in the past (what we call statistical forecasting) and informed judgments about future events [...]" (Silver et al., 2016, p. 73). Most companies consider demand a purely time-dependent variable,

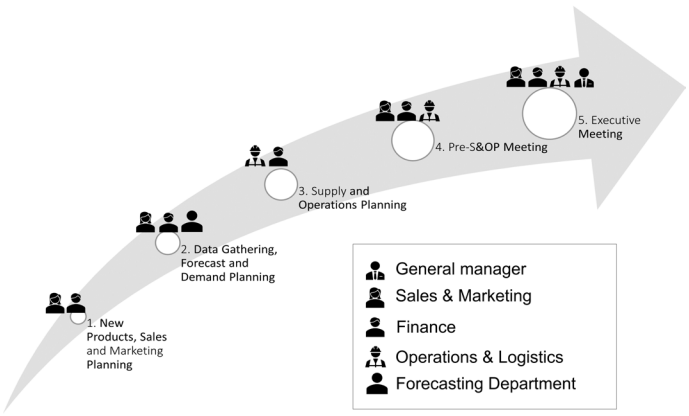

Figure 1. S&OP Process. Source: (Castro-Zuluaga & Arboleda-Florez, 2021).

so statistical forecasting focuses on one-dimensional time-series models (Brockwell & Davis, 1987; Chatfield, 2000). Based on experts' judgments, such forecasts are later adjusted through human input. For example, the marketing department could provide the impacts of promotions, upcoming trends, and expected moves from competitors to make a positive or negative adjustment to the statistical forecasting obtained using any model. Other areas could include the effects of preexisting orders, special economic conditions, or new product releases, impacting the final forecast.

This approach has remained acceptable for decades for various reasons: 1) Their univariate nature makes traditional time-series models easily interpretable. Therefore, analysts can create 2-D charts for time-series models to visualize demand changes (y-axis) through time (x-axis). Visualization provides a simple form to inspect each product's behavior visually. Analysts can then identify stationarity, seasonality, cyclicality, and trends, annotate abnormal periods, and compare results for different time windows. 2) Different people can contribute while keeping a narrow perspective. Each expert can focus on forecasting the expected effect of their strategies, analyzing a few variables at a time and inside their area of expertise. Thus, the responsibility of having a good-enough forecast is distributed across multiple people working together to reach an agreement while the forecast remains univariate. The next S&OP steps receive and analyze a single report with charts and annotations without worrying about the internals of those results. A simple model does not mean that the forecasting process is simple but that the complexity is placed on the human input and judgment instead of the modeling process, which makes it more prone to errors.

Nevertheless, as stated by Makridakis (1988, p. 475), "[...] intuitive or judgmental forecasts can bring large and systematic errors caused by biases in the way information about the future is recalled and processed." People's optimism and wishful thinking could lead to overestimation, while their beliefs and backgrounds could create illusory correlations and inconsistencies. For example, personal preferences around specific products or strategies could increase expectations about their future performance. At the same time, a negative bias could lead to underestimating other products. Human judgment is difficult to measure and replicate, so including it in this forecasting method fails to answer critical questions like how much of the forecasting error was caused by the model and how much of it was caused by human judgment, or even how different variables affected different products.

Now, in the era of data and automation, new solutions have emerged to reduce uncertainty. Factors such as the market's globalization and technological innovation have added complexity to every company (Vogel & Lasch, 2016). Nevertheless, they have also provided the systems to collect, process, and analyze data. Companies can gather enough data about their products, customers, and competitors, integrate multiple sources of information, compute complex calculations, and test several hypotheses at a low cost and short time. This amount of data has fueled machine learning methods as a solution to improve accuracy, optimize processing time, and reduce human interaction in a wide range of applications (Yao et al., 2018).

Machine learning refers to the set of methods for discovering patterns in data without being explicitly programmed (Bisong, 2019). Machine learning algorithms are of significant use in "making sense of data," which means finding and learning from patterns in data to make predictions, uncover hidden structures, and find meaningful relationships among data (Raschka & Mirjalili, 2015). These algorithms already play an essential role in applications proven too complicated for traditional programming approaches, such as face and speech

recognition problems (Mitchell, 1997). They also have outperformed alternative methods in supply chain applications (e.g., customer segmentation. Hiziroglu, 2013), supply chain risk prediction, inventory management (Tirkolaee et al., 2021), and demand forecasting (Carbonneau et al., 2008).

However, most machine learning models are black-box models because of their high complexity and low interpretability. It is challenging to develop a comprehensive understanding and ensure trust in their predictions (Shams Amiri et al., 2021). The latter has become a breaking point for implementing machine learning models beyond academic research. For example, Ishikawa & Yoshioka (2019) surveyed the perceived difficulties when implementing machine learning systems. They found that current practitioners mention that interaction with internal customers is the most challenging part of the process, mainly because customers lack an understanding of how the models work and reject the uncertainty in the relationships between input and output. Likewise, when people in managerial positions were asked about their concerns about implementing machine learning in their companies, one of the critical points they mentioned was the lack of explanation for the output of the created models (Gartner Inc., 2017). Which agreed with the results obtained by Deloitte (2017) that 47% of surveyed industry leaders think it is challenging to integrate machine learning projects with existing processes, and 37% of them think that one of the implementation challenges is that managers do not understand cognitive technologies and how they work.

Explainable Artificial Intelligence (XAI) has emerged as a solution to this problem. As stated by (Barredo Arrieta et al., 2020, p. 83), "XAI proposes creating a suite of ML techniques that produce more explainable models while 1) maintaining a high level of learning performance (e.g., prediction accuracy), and 2) enabling humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners". This paper focuses on using the recently proposed SHAP (SHapley Additive exPlanations) method as a resource to explain the decision process behind a forecasting model for one direct sales company and to encourage the transition from traditional to automated forecasting techniques. The SHAP method provides a robust theoretical basis for interpreting prediction models by representing each feature's influence on the model output while ensuring feature consistency and model stability (Meng et al., 2020). By proposing the inclusion of SHAP in the automation of the forecasting process, we aim to reduce friction across the S&OP process against implementing machine learning algorithms. Achieving interpretable visualizations could minimize the project's failure risk, augment confidence in the model and lead to full adoption of the new technology, ultimately increasing the company's resilience by adapting to technological innovations and dealing with the increased demand uncertainty.

In this paper, we describe the suggested methodology to include both the SHAP method and human knowledge into a machine learning pipeline to reduce the risk of its implementation in a direct sales company's forecasting system. The rest of the article is as follows. Section 2 presents the relevant background referring to Machine learning models in demand forecasting and the SHAP method. Section 3 describes the company, its present state, and the data. Section 4 describes the proposed methodology for implementing machine learning models in the company. In section 5, we show implementation results, and in section 5, we present a conclusion and future work. The main emphasis is laid on the flexibility and versatility of implementing SHAP as a tool to interpret and analyze the results of automated forecasting using machine learning techniques.

2. Background

2.1. Machine learning

Machine learning refers to methods that computers use to make and improve predictions or behaviors based on data (Sculley et al., 2014). We say that a machine learns "[...] concerning a particular task (T), performance metric (P), and type of experience (E), if the system reliably improves its performance (P) at the task (T)" (Mitchell, 2006, p. 1). In other words, Machine learning refers to algorithms that gradually improve their accuracy by automatically adjusting their parameters based on an error function and the relationships learned from examples. The ability to learn differs from traditional programming in software engineering, where rules and conditionals define the algorithm's behavior that human engineers explicitly program based on their knowledge.

Machine learning is considered a "[...] general-purpose technology like the steam engine and electricity [...]" (Brynjolfsson & Mitchell, 2017, p. 1530), which can be used in multiple applications and innovations. Machine learning algorithms divide into supervised, unsupervised, or reinforcement learning depending on the type of data available and the task it is expected to solve. Supervised learning refers to those algorithms that learn from labeled data. The algorithms receive the features and the actual output for each past observation. The algorithm then iterates and proposes a model that optimizes a loss function, minimizing the difference between the

predicted and the actual value. Supervised learning algorithms are commonly used for predictive tasks such as classification problems and regression problems, for instance, to predict customer churn (Vafeiadis et al., 2015) or product quality (Ktenioudaki et al., 2021). Unsupervised learning refers to algorithms that receive unlabeled data and are expected to recognize patterns and similarities across data. These algorithms are used as an exploratory step where grouping data is essential, for example, identifying categories or clustering similar customers to select different strategies (Sheshasaayee & Logeshwari, 2018). Lastly, reinforcement learning refers to algorithms that work better with interactive problems in which it is essential to sense the state of the environment and take action accordingly (Sutton & Barto, 2018). There is a reward for sensible approaches and a penalty for wrong actions (Wenzel et al., 2019). Reinforcement algorithms have predominantly been used in robotics (Kormushev et al., 2013) and transport management (Abdulhai & Kattan, 2003).

Accordingly, demand forecasting is, in most cases, a supervised regression problem. Companies have historical data on their products' characteristics, marketing strategies, related external factors (input variables), and demand (expected output). Therefore, it is possible to train and evaluate models iteratively in search of the optimal solution. Numerous studies have investigated machine learning implementations in sales forecasting. Some of the most popular implemented algorithms are tree-based machine learning algorithms and neural networks. For instance, Krishna et al. (2018) and Dairu & Shilong (2021) used gradient-boosted trees to predict sales for retail stores. Similarly, Gumani et al. (2017) compared and combined various methods on one drug store data and found that gradient-boosting trees best capture non-linear components and trends. Regarding the fashion and apparel industry, there has been an increasing interest in combining methods and implementing neural networks (Chen & Lu, 2021; Lorente-Leyva et al., 2021; Ren et al., 2020; Sun et al., 2008).

Existing comparisons between machine learning and traditional time series models have shown that machine learning models accomplish higher accuracy and additional flexibility to handle a significant number of data variables, unclear tendencies, and intermittency between sales (Jeon & Seong, 2021; Seaman & Bowman, 2021). Machine learning solutions outperformed traditional methods when performing complex analyses, such as the impact of promotions (Tarallo et al., 2019). However, despite the proven capacities machine learning-based solutions have shown, they still face challenges that hinder their deployment in more traditional, change-adverse organizations, mainly because of their lack of transparency and interpretability (Samek & Müller, 2019). Science has turned to Explainable Artificial Intelligence (XAI) techniques that produce more interpretable outcomes while maintaining the predicting power of Machine Learning solutions.

2.2. SHAP values

Adadi & Berrada (2018, p. 52139) define Explainable Artificial Intelligence (XAI) as a "[...] research field that aims to make AI systems results more understandable to humans." However, there is not yet a standard definition of XAI, which is also defined with words such as transparency, interpretability, and explainability (Clinciu & Hastie, 2019). While the interest in developing such solutions has radically increased in recent years, early investigations studied explanations for expert systems. They recognized that explaining a system's results was critical to the intended users, who perceived more importance in explanations than in having "perfect" results (Moore & Swartout, 1988).

There are different approaches to explainable AI. Some of the most common techniques are model agnostic, which means that the technique can be applied to "[...] any classifier, even without knowing its internals, e.g., architecture or weights of a neural network classifier [...]" (Samek & Müller, 2019, p.12). Model-agnostic XAI techniques involve perturbing inputs of the original model and identifying their impact on the prediction (Strumbelj & Kononenko, 2010). Model-agnosticism provides model flexibility and homogenous comparison enabling the use and evaluation of any machine-learning model (Ribeiro et al., 2016). Comparison helps the company define and maintain one transparent communication system across all products, even when they are forecasted using different models.

This study focuses on implementing SHAP (SHapley Additive exPlanations), one of the most popular model-agnostic techniques for explainable AI. SHAP is a model additive explanation approach from cooperative game theory. The method presents and explains the prediction concerning the contribution of each feature to the predicted value (Bugaj et al., 2021). Being a model-agnostic methodology, SHAP can explain individual predictions without being limited to a specific machine-learning model. SHAP considers the uniqueness of each prediction while highlighting the global factors influencing the overall performance (Kumar & Boulanger, 2020). In other words, SHAP uses game theory to evaluate each model's prediction "as a model itself" (a game) and treats every input feature as a player for each game. Just as players can participate in a specific game, each feature can join or not join a model. Therefore, only present features will contribute to the model's outcome. The explanations are then built by asking how prediction p changes when feature f is removed from the model (Gilbert, 2019), and SHAP values answer those questions.

SHAP values are, therefore, the measure of feature importance and adhere to three desirable properties that increase interpretability in different scenarios. These properties are 1) Local accuracy, where the explanation model should match the output of the original model, 2) Missingness, meaning that whenever a particular variable is missing in the original input, it should have no impact on the explanation; and 3) Consistency, which states that if in a different model the contribution of a variable is higher, so should be the corresponding Shapley value (Lundberg & Lee, 2017). Gramegna & Giudici (2021) mentioned that other popular methods violate at least one of the properties. Also, the SHAP Python package is strongly supported by an extensive community continuously working towards its improvement. Such a powerful package fitting the Machine Learning pipeline at no additional cost is advantageous and secures long-term project survival.

A key characteristic of SHAP values is that they measure the magnitude and direction of a feature's effect on a prediction. Measuring both properties of the effect allows the sum of all contributions to equal the difference between the baseline model output and the current model output. SHAP's characteristics have grown its acceptance and forged a way to become a widely accepted unified measure of features' importance.

3. Problem definition

This research focused on the sales forecasting process from a direct sales company dedicated to producing and selling apparel goods. Sales forecasting is a complex process that almost any company must face. This process faces unique difficulties in the direct sales industry regarding its selling model. For instance, direct sales companies sell through printed catalogs that their independent representatives distribute and promote to collect and place unified orders later. Because of this, companies perceive their demand in discrete-time intervals instead of a continuous flow. Also, their catalog differs for each campaign, meaning that products receive different exposition levels every time. Moreover, catalogs require photos of each product, visual design, printing, and distribution, increasing the overall lead time beyond the manufacturing process, which increases the process' leading time, and extends the forecasting window and the probability of erring (Castro-Zuluaga & Arboleda, 2019).

In the past, the company focused on one segment of products with few well-known references displayed in their catalog. Since then, the demand forecasting process has relied on a team to plan for the upcoming months through traditional time series models and judgmental methods. Nevertheless, the company has amplified its variety of products over the past few years and expanded its presence across different cities. Their expansion has increased the difficulty of providing on-time and accurate forecasts, so the process has become slow, unreproducible, and prone to errors. Furthermore, data has become too complicated to analyze individually, leading to growing tension and lacking trust between functional areas.

Automating and optimizing the forecasting process would save time and money while positively affecting the supply chain. However, most people along the supply chain are change-averse and disregard most attempts to introduce new forecasting methods, especially those they do not understand. As was mentioned above, we found in SHAP a way to use as an explanatory tool for managers and stakeholders outside the forecasting area to understand and approve the demand forecasts that the model proposes. SHAP's visualizations are intuitive and easy to explain at multiple levels of detail. They could be used for each step of the S&OP process and engage all areas in meaningful conversations around the proposed forecast.

4. Methodology

This empirical research is of a quantitative-explanatory type, and the method used is inductive and based on observations (Bertrand, & Fransoo, 2002). This study aims to illustrate how methodologies like SHAP may facilitate adopting an automated forecasting process by explaining an otherwise black-box model to restructure the company's forecasting process.

The traditional forecasting process used by the company, as shown in Figure 2, starts by using a statistical model to forecast future sales solely based on historical sales data (sales at each time window). Different experts review the statistical forecast in a multidisciplinary meeting and provide their knowledge to adjust it. Then, the adjusted forecast is used throughout the rest of the S&OP process. Here, human judgment is seen as a secondary step to improve the model's accuracy, but, as the human thinking process is impossible to standardize, it lacks objective feedback and a replicable measure of error. One way to think about this is to ask: if three people adjusted the same forecasted quantity based on their intuition, how much error, if any, was introduced by each person? Moreover, would the same people make the same decisions if the same data were presented to them later? Probably, the answer would be no, because decisions are not standard, but machines and algorithms are.

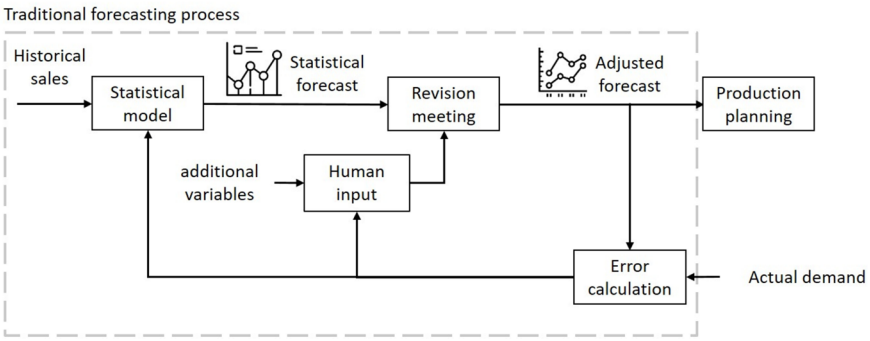

Figure 2. Traditional forecasting process.

Implementing machine learning methods incorporates human judgment in the planning and modeling process. That way, experts' judgment is expected to feed the machine learning model through data preprocessing, parameter tuning, and providing additional structured information (e.g., managerial decisions). Figure 3 presents the proposed forecasting process where human input has moved a step back to support the data and model preparation process by providing crucial business knowledge. The rest of the process is automatized and therefore documented and replicable.

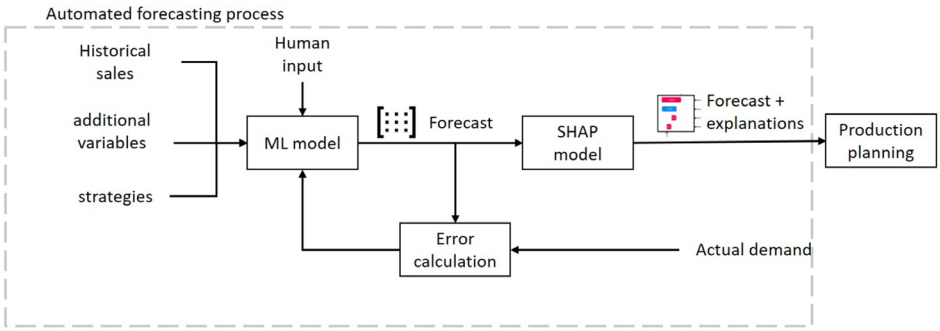

Figure 3. Proposed forecasting process.

The proposed methodology consists of five main steps: data collection, preprocessing, feature engineering, model selection, and model explanation. The last step concludes with integrating SHAP in the machine learning pipeline to extract valuable information about how the model interprets the data and the reasons behind its results. The information clearly communicates the results to the intended users beyond the forecasting department.

This study was designed for a direct sales company. The company creates approximately 18 catalogs annually, with more than 4000 products ranging from primary home products to apparel items. Those catalogs are distributed across the country and promoted by independent representatives who later collect and unify all their customers' orders. The independent representatives are also responsible for distributing the products to the final customers and collecting their payments. For this project's scope, we analyzed the apparel items' sales, marketing, and design data of almost 90 catalogs.

The first step, *data collection*, involved collecting databases scattered across different company functional areas. For instance, while the forecasting department provided historical sales data, it was necessary to ask the marketing, design, and production departments to share their databases to integrate their decisions and knowledge into the analysis. Most company areas work with personalized excel files where the information is grouped and presented according to their needs. Therefore, this step required disaggregating and translating the information into product-related features. For instance, the duration of a campaign, seasonal sales, and many special events affect all products sold during that time. Thus, we performed one-to-many merges to assign each event to all necessary items and obtain one tabular dataset, as presented in Figure 4. The dataset contained five years of historical data and 12 selected features. The dataset included product-related variables, i.e., size, color, and price; marketing decisions, such as a promotion or publicity during a campaign; and time-related variables, like the days each catalog was available to the public.

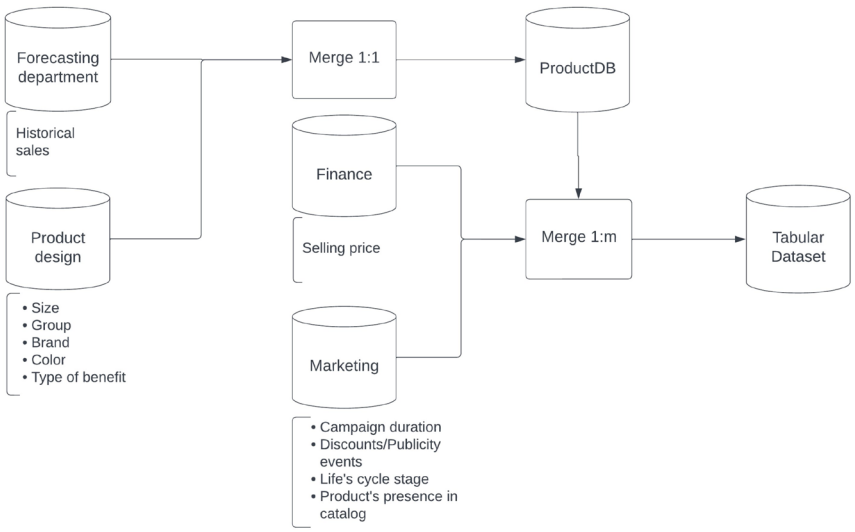

Figure 4. Data collection.

Most of the data is categorical and manually entered during each step of the design and production process. The *data preprocessing step* required unifying categories, identifying and imputing missing values, and deleting duplicated data. Unifying categories meant that we selected the most prominent color of each item and defined a standardized name to use. This way, we replaced various shades, names, and codes that different employees used for the same color with the new standard (for example, we turned navy, dark blue, and light blue into one category called blue). With the help of business experts, we also identified some company-specific categories (and exceptions), such as *patterned* and *denim*. Similarly, we translated marketing strategies into the effect perceived by the customer, e.g, turning "Christmas Sale" into a standard category called "Discount"; reducing more than 40 categories with scattered data to only 14.

Regarding data imputation for missing values, we inspected each feature to decide which method would be the most meaningful. First, we identified that a missing value for marketing strategies and events means that the product did not have any promotion or strategy attached during the campaign, so we created a new "No Event" category. Then, we used the most common value inside the catalog and product category for imputing missing decision-based categorical values such as the event and life cycle. For product-based categories, e.g., size, benefit, and brand, we imputed the last available value for similar products in the previous campaigns; lastly, we used the mean value of similar products for imputing missing price values. Table 1 describes some of the main characteristics of the products with examples of their values.

Human input or expert judgment was crucial during data gathering and preprocessing. As was presented in Figure 3, we aimed to "move" the human input a step back, and we did so by 1) collecting and integrating information to the dataset that would otherwise be processed, analyzed, and presented individually; 2) including the business experts into the decision making process of reducing the number of levels for each feature, generating new categories, understanding the difference and purpose of multiple variables and how to approach them; and 3) as we discuss further in the results, by providing meaningful insights on what to look for when evaluating the results.

Table 1. Description of products' features.

FEATURE	DESCRIPTION	VALUES
SIZE	Each of the classes into which garments are divided	XS-XXL, 6 - 14, 30 -36
EVENT	Type of marketing event a product can be part of	Publicity, discount, extra product, no event
LIFE_CYCLE	Stage that the product is in	Introduction, growth, maturity, decline
BRAND	Classification according to the public a product is aimed at	Women's, men's, juvenile
GROUP	Type of item	Shirt, skirt, pants, ...
COLOR	Main color of the product	Yellow, blue, black, indigo, ...
PRICE	Lower limit of the range that contains the product's price	20000, 30000, ..., 90000, 100000+
HAS_FEATURE	Type of special feature or benefit a product may have	Thick cloth, breathability, ...

Feature generation increases the models' predictive force and overall performance (Dong & Liu, 2018). Feature generation was of particular value in the project as it introduced additional numerical data and the direct relationship between time and categorical features. We generated two groups of features: moving averages of the last five periods' demand and the total demand in the same campaign year before for each category. As a result, the dataset contains two variables for each category (price, color, size, benefit, event, and stage) that describe each product by its features' history.

The dataset consists of more than 57000 observations, 36 features, and a target variable, saved in a CSV format that follows the structure shown in Table 2, and a dictionary stored in a JSON format containing the types and column names associated with the data set, as presented in Figure 5. It is important to note that the variable *Demand* is the actual historical data, against which we will evaluate our models. This resulting structure is considered the starting point of the proposed forecasting process and must be preserved during the upcoming stages and future experiments.

Table 2. Dataset structure.

Item	Campaign	Feature_1	Feature_2	...	Feature n	Demand
1	C_1					y_1
2	C_1					y_2
3	C_1					y_3
.	...					.
.	C_2					.
.	...					.
m-1	...					y_(m-1)
m	C_j					y_m

```
{
   "categorical_cols":
   [
      "Campaign",
      "Feature_1",
      "Feature_2",
      ...,
      "Feature_x"
   ],
   "numerical_cols":
   [
      "Demand",
      "Feature_k",
      "Feature_n"
   ]
}
```

Figure 5. Dictionary of data structure.

The next step of the process involved *the selection of the machine learning model* to predict future sales for each item based on the information provided. As mentioned before, one of our limitations is that the company relied on expert judgment to forecast, which meant there was no reliable baseline to compare. Consequently, we used a simple naïve model as a baseline for model performance to evaluate the best machine learning approach. Then, we separated the testing data from the training set, and only then we proceeded to train and evaluate the four machine-learning algorithms. The tested algorithms range from one of the simplest models to one state-of-the-art gradient-boosting tree-based model. The machine learning pipeline was programmed using Python 3.8 on a personal computer with 16Gb Ram. All the models we tested are available in the Scikit-learn package (Pedregosa et al., 2011), except for Catboost, for which we used their Python package. We present a brief explanation of each tested model:

- Baseline: Refers to the naïve approach, which predicts that each product will continue to sell the average of its past sales. As the company had worked solely with intuitive forecasting, it was essential to set a replicable and consistent baseline to compare the effectiveness of the machine learning models.
- Linear: Refers to a linear regression model. It models the regression function as a linear combination of predictors, and the model parameters are easily interpretable (Su et al., 2012).
- Linear reg: Refers to a regularized linear regression model that penalizes complex methods and aims to control overfitting. (Friedman & Popescu, 2003)
- Random forest: Refers to a popular tree-based ensemble model that, in the words of Breiman (2001, p. 5), "[...] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [...]" and when after a large number of trees is generated, "[...] the random forest predictor is formed by taking the average over k of the trees [...]."
- Catboost: Refers to a robust machine-learning algorithm that uses gradient boosting on decision trees (Prokhorenkova et al., 2019). Catboost also outperformed XGBoost and LightGBM while working with categorical features using combined category features to enrich feature dimensions (Zhang & Ma, 2020)

Figure 6 presents the distribution of the absolute error of each model. The box length represents 50% of errors, and the length of the whiskers represents the bottom and top 25% of errors. In our case, we observed that the baseline model had the most variability and the highest error in general, having its median around 60 and an interquartile range of 40. Both linear models performed similarly and better than the baseline. The median error is significantly lower, but the variability remains high. Finally, random forest and catboost presented the lowest errors with the lowest variability, as shown by the small box and short whiskers.

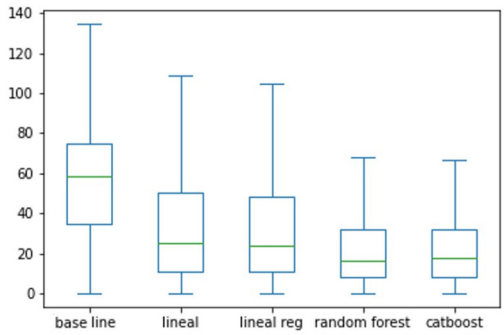

Figure 6. Absolute error for tested models.

The figure shows how machine learning approaches were significantly more accurate than the baseline approach, with tree-based algorithms performing better than the simpler linear models, being more accurate and having less variability. CatBoost was selected for its performance and optimization to work with multiple categorical features from the dataset. Also, CatBoost's python package integrates SHAP based on the Consistent Individualized Feature Attribution for Tree Ensembles approach (Lundberg et al., 2019). It does not require further data transformation, such as categorical encoding or feature standardization.

After choosing the best-fitting model, we retrained the Catboost regression model considering all the available data and moved to the last step of the process. A fully functional demo of the proposed methodology is available at https://mariaf494.github.io/explain-forecast/. It is possible to interact with the results, see different plots than those presented in the article, and load different datasets to test the methodology.

The last step consisted of computing SHAP values to *explain the model.* We generated the explaining visualizations using the public python package (Lundberg, 2019). Until this point, the chosen forecasting model remained a black box. However, SHAP values give accurate information about which feature contributed to the predicted value and an estimate of its overall impact on the prediction. We provide four types of visualizations to understand the process, which will support the analysis of the obtained forecast throughout the following stages of the S&OP.

5. Results and discussion

Our model aims to predict each product's demand given its features. The initial output is an array of numbers that is not meaningful for the company for a given set of products. However, when we plot our model's SHAP values, it provides meaningful insights into the outcome.

As stated in the previous section, the SHAP values evaluate each feature's contribution to each prediction. It is important to note that the base value corresponds to "[...] the value that would be predicted if we did not know any features for the current output [...]" (Lundberg & Lee, 2017, p. 5) and is set to the average demand of the corresponding dataset. Feature contribution refers to how that feature's presence explains the difference between the actual forecast and the base value. The SHAP visualizations illustrate the contribution by using colored labels. Red bars and points refer to positive contributions that increase the forecasted value beyond the base value, while blue bars and points refer to negative contributions that decrease the forecasted value below the base value.

Force plots illustrate the interaction between all features and their contribution to individual predictions. Figure 7 presents an example of four products and how the forces of their features interact to "push" their forecasted demand higher (forces in red) or lower (forces in blue). For all cases, the base value is around 179 units, meaning that if there were no information about any of the products, the model's best guess would be to forecast 179 units for all products. For product (a), the model forecasted 143 units. When looking into detail, it is possible to observe that the price, size, and historical information of its event, group, and color, all contribute negatively to the prediction. A negative contribution means that the products' characteristics tend to perform worse than the expected value.

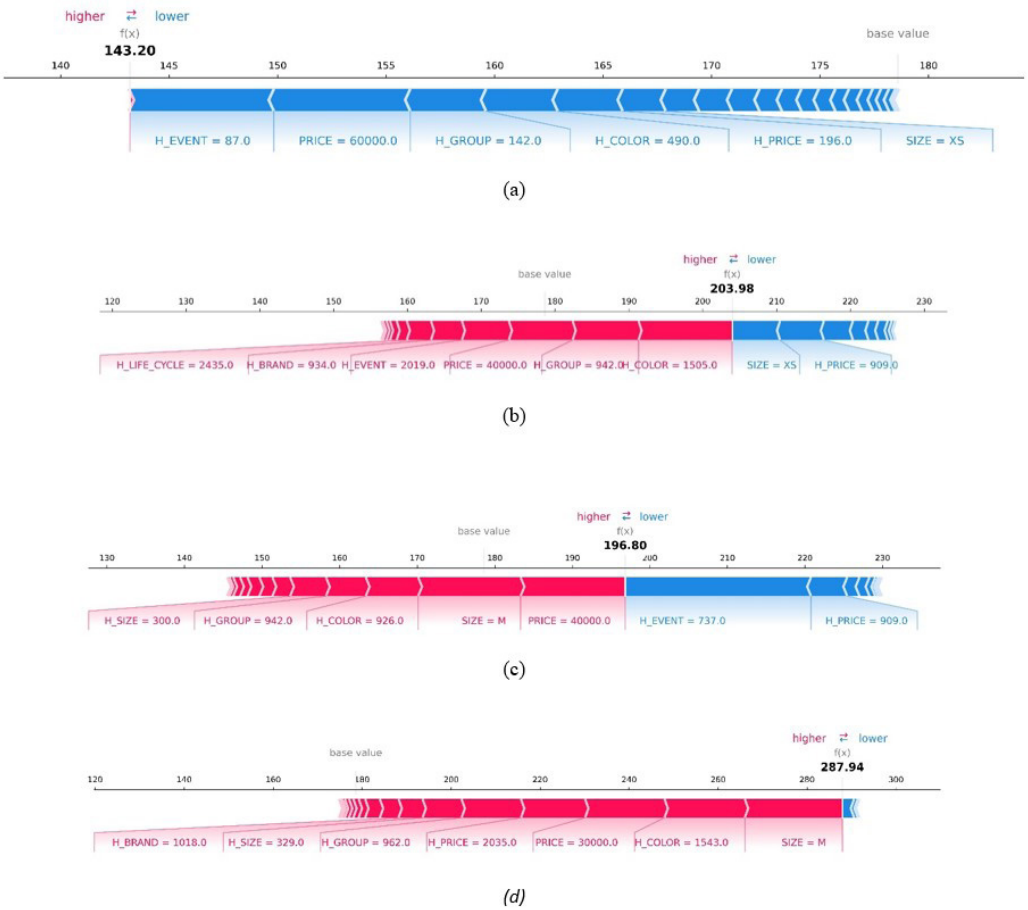

Figure 7. Force plots.

On the contrary, product (d) presents mostly positive interactions associated with its size, color, and price, which explains the predicted value of 288, almost 100 units above the baseline. Products (b) and (c) present negative and positive contributions, given their features. It is important to note that the *direction* of one level of a feature is consistent across the different products. For example, products (a) and (b) are both size XS, while products (c) and (d) are size M. A product being of size XS contributed negatively to the forecast as opposed to being size M. This means that the company can expect to sell more Medium-size products than it is expected to sell extra small. The difference in the contribution of both sizes is coherent with what experts had observed through experience: the medium size is the most popular for this context, as it is designed to fit the customers' mean size. We also observe a similar effect in the feature Price. Of the four products shown, product (a) is the most expensive and the only one where the price contributes negatively to the forecast. For products (b), (c), and (d), having a price lower than COP 45.000 (around USD 11,5) contributes positively to the prediction. Again, the company adheres to the observed effects, setting a precedent to turn opinions into factual inferences.

The waterfall plot shown in Figure 8 provides an alternative visualization to the force plots mentioned above. This way, it is possible to zoom into the magnitude of each effect and follow the model's prediction. Considering that the company has the detailed product description, the company can contrast these visualizations to the products' description, group and analyze the results according to the company's needs. For instance, in cases where there is a new product release with high expectations of its performance but the predicted value does not correspond, analyzing its waterfall plot could provide a better idea of the interactions that lead to the forecast.

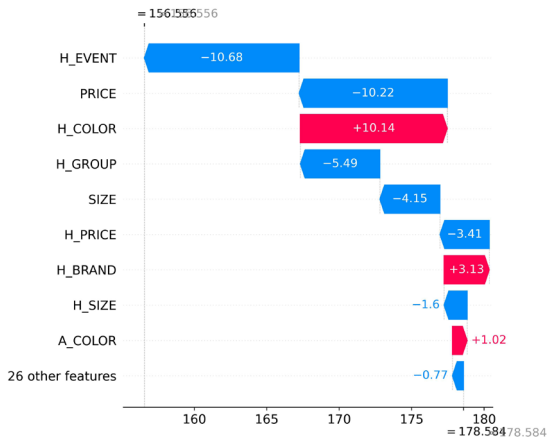

Figure 8. Waterfall plot.

Now, on a global manner, Figure 9 summarizes the impact of each feature on our model. The figure presents the sorted features according to their average impact, and only the 20 most important features are shown. So, in this case, one of the generated features representing the item's historical performance (H_EVENT) is the most critical feature for predictions. This figure also allows us to understand which information is vital for the model, and although it is out of the scope of this paper, it would be an essential input for performing feature selection (Marcilio & Eler, 2020).

Finally, having this resource, it is possible to gain essential insights into the non-linear relations usually hidden in data at various levels of detail. Through the analysis of the different force plots, we concluded that size M always had a positive impact on predictions, while other sizes had negative impacts. Figure 10a supports this conclusion by providing a graphical summary of each size's impact on all our observations. In this figure, the feature size M has the most significant positive effect on predictions, followed by sizes S and L, which also present positive impacts. It is evident that the relationship is non-linear and cannot be captured by a traditional time-series model.

Similarly, Figure 10b presents the price's impact on the forecast. The visualization shows that even though more affordable products have positive contributions to the model while more expensive ones have negative contributions, the expected contribution does not follow a linear manner. The figure reveals the frontier between affordable and expensive products in the company's context. It is essential to mention that the specific value obtained for this data may be meaningless when translating to other companies, even for ones with similar conditions, but the methodology impacts the understanding of the business at all managerial levels.

In conclusion, the SHAP method allows us to "open the black box" and provide enough information about the model's output, addressing the challenge mentioned by Gartner Inc. (2017) without requiring the audience to understand the inner working of the machine learning models.

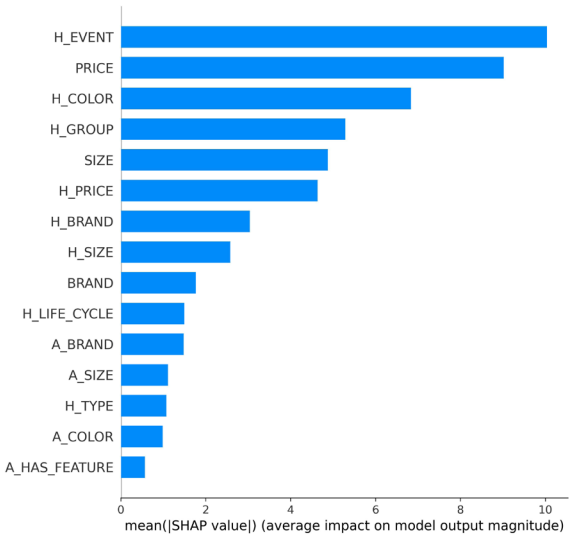

Figure 9. Summary plot.

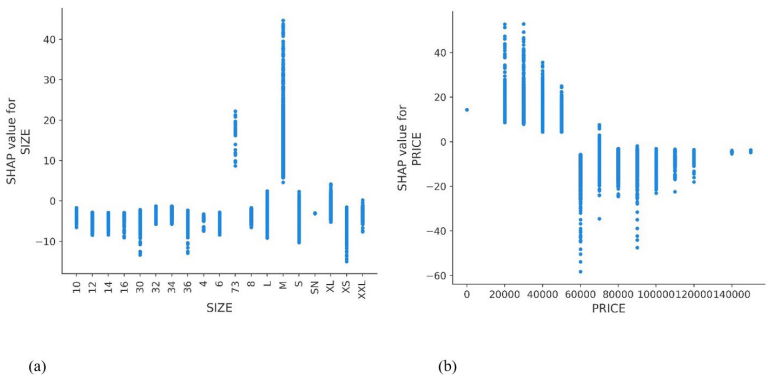

Figure 10. Dependence plot.

SHAP implementation addresses the two main challenges identified in the surveys conducted by Deloitte (2017): Managers understanding the technologies and integrating machine learning within the existing process the company can start. As a result, the company can start building trust in the machine-learning approach and benefit from its implementation.

In the early stages of the implementations, the main expected benefits are:

- To have the flexibility to work with many variables and perform complex analyses.
- To provide a better understanding of the effect of different strategies and product characteristics, as well as their interactions with one another, compared to the understanding achievable by human reasoning alone.
- To reduce the time invested in forecasting: since the data preprocessing and forecasting are fully automated, the efforts concentrate on results analysis.
- To estimate a standard measure of error and a replicable process for the company to conduct error analysis. Therefore, the company can inspect the forecasting errors and focus on products performing worse than others.
- To perform experiments to improve accuracy without affecting the model in operation. The described machine learning pipeline is prone to modifications and continuous improvement. The company can test different hypotheses without them reflecting on cost for the company.

In the long run, the implementation would lead to more significant benefits, such as increasing the level of precision in demand forecasting and improving inventory management and efficiency; "[...] thereby resulting in cost savings, increased revenue, and greater customer satisfaction [...]" (Tarallo et al., 2019, p. 741).

6. Conclusion

Demand forecasting is a crucial task in the Sales and Operations Planning Process. Globalization, technological developments, increasingly demanding customers, and market conditions have added complexity to all companies. Technological development has also allowed companies to generate and collect large amounts of data regarding their products, clients, and processes, setting the terrain to exploit the strengths of machine learning. However, due to the lack of trust, there is the black-box nature of most machine learning models; those companies still rely on traditional methods and human judgment to forecast their demand, evidencing low accuracy with significant time investment.

This study presented the introduction of SHapley Additive exPlanations (SHAP) as part of the automated forecasting process of a direct sales company. The proposed methodology aims to reduce the friction of implementing new technologies as part of the S&OP process by allowing the company to visually understand the predictions made by an otherwise black-box model. It is relevant to note that the methodology restructures the existing process but does not replace the human judgment or expertise of the company's functional areas. Results provide important insights regarding the complex multivariate interactions that influence the demand and serve as a verification method to contrast intuitions and ease the transition to data-based decisions.

The methodology also sets the foundation to continue working towards adopting new technologies to improve the company's resilience. This research used exclusively structured data already available inside the company; future work could integrate unstructured data, such as the pictures used on catalogs, and add external variables to strengthen the analysis.

References

Abdulhai, B., & Kattan, L. (2003). Reinforcement learning: introduction to theory and potential for transport applications. *Canadian Journal of Civil Engineering*, *30*(6), 981-991. http://dx.doi.org/10.1139/l03-014.

Adadi, A., & Berrada, M. (2018). Peeking inside the black box: a survey on explainable artificial intelligence (XAI). In *Proceedings of the IEEE Access: Practical Innovations, Open Solutions* (Vol. *6*, pp. 52138-52160). USA: IEEE. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

Bandeira, S., Alcalá, S., Vita, R., & Barbosa, T. (2020). Comparison of selection and combination strategies for demand forecasting methods. *Production*, *30*, e20200009. http://dx.doi.org/10.1590/0103-6513.20200009.

Barredo Arrieta, A., Díaz-Rodríguez, N., del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. http://dx.doi.org/10.1016/j.inffus.2019.12.012.

Bertrand, J. W. M., & Fransoo, J. C. (2002). Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, *22*(2), 241-264.

Bisong, E. (2019). What is machine learning? In E. Bisong. *Building machine learning and deep learning models on google cloud platform* (pp. 169-170). Berkeley: Apress. http://dx.doi.org/10.1007/978-1-4842-4470-8_13.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324.

Brockwell, P. J., & Davis, R. A. (1987). *Time series: theory and methods*. New York: Springer. http://dx.doi.org/10.1007/978-1-4899-0004-3

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534. http://dx.doi.org/10.1126/science.aap8062. PMid:29269459.

Bugaj, M., Wrobel, K., & Iwaniec, J. (2021, May 12-16). Model explainability using SHAP values for LightGBM predictions. In *International Conference on Perspective Technologies and Methods in MEMS Design*. USA: IEEE. http://dx.doi.org/10.1109/MEMSTECH53091.2021.9468078.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140-1154. http://dx.doi.org/10.1016/j.ejor.2006.12.004.

Castro-Zuluaga, C. A., & Arboleda, M. (2019). Sales forecasting difficulties' analysis on colombian direct sales companies. In M. T. Castañeda Galvis, J. Nuñez Rodriguez, M. C. Pérez Ordoñez & M. Villa Marulanda (Eds.), *Proceedings of the International Congress of Industrial Engineering (ICIE2019). ICIE 2019. Lecture Notes on Multidisciplinary Industrial Engineering* (pp. 112-118). Cham: Springer. https://doi.org/10.1007/978-3-030-49370-7_12.

Castro-Zuluaga, C., & Arboleda-Florez, M. (2021). Introduction. In M. Hemmati & M. S. Sajadieh (Eds.), *Influencing customer demand: an operations management approach* (pp. 1-16). Boca Raton: CRC Press. http://dx.doi.org/10.1201/9781003107446-1.

Chatfield, C. (2000). *Time-series forecasting*. London: Chapman and Hall/CRC. https://doi.org/10.1201/9781420036206.

Chen, I. F., & Lu, C. J. (2021). Demand forecasting for multichannel fashion retailers by integrating clustering and machine learning algorithms. *Processes*, *9*(9), 1578. https://doi.org/10.3390/PR9091578.

Clinciu, M. A., & Hastie, H. (2019). A survey of explainable AI terminology. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)* (pp. 8-13). http://dx.doi.org/10.18653/v1/W19-8403.

Crum, C., & Palmatier, G. E. (2003). *Demand management best practices: process, principles, and collaboration*. Boca Raton: J. Ross Publishing.

Dairu, X., & Shilong, Z. (2021). Machine learning model for sales forecasting by using XGBoost. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering, ICCECE 2021* (pp. 480-483). USA: IEEE. http://dx.doi.org/10.1109/ICCECE51280.2021.9342304.

Deloitte. (2017). *Bullish on the business value of cognitive Leaders in cognitive and AI weigh in on what's working and what's next*. New York: Deloitte.

Dong, G., & Liu, H. (Eds.). (2018). *Feature engineering for machine learning and data analytics*. Boca Raton: CRC Press.

Friedman, J., & Popescu, B. E. (2003). Gradient directed regularization for linear regression and classification. Technical Report, Statistics Department, Stanford University.

Gartner Inc. (2017). *Analysts Answer What Are Top Client Concerns About Ai*. Retrieved January 24, 2022, from https://www.gartner.com/smarterwithgartner/analysts-answer-what-are-top-client-concerns-about-ai

Gilbert, F. (2019). *Introducing SHAP Decision Plots. Visualize the inner workings of machine learning models with greater detail and flexibility*. Towards Data Science. Retrieved January 24, 2022, from https://towardsdatascience.com/introducing-shap-decision-plots-52ed3b4a1cba

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558. http://dx.doi.org/10.3389/frai.2021.752558. PMid:34604738.

Gumani, M., Korke, Y., Shah, P., Udmale, S., Sambhe, V., & Bhirud, S. (2017). Forecasting of sales by using fusion of machine learning techniques. In *2017 International Conference on Data Management, Analytics and Innovation, ICDMAI 2017*, (pp. 93-101). USA: IEEE. https://doi.org/10.1109/ICDMAI.2017.8073492.

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: state-of-art review and critique. *Expert Systems with Applications*, *40*(16), 6491-6507. http://dx.doi.org/10.1016/j.eswa.2013.05.052.

Ishikawa, F., & Yoshioka, N. (2019). How Do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems? - Questionnaire Survey. In *Proceedings of the 2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP)* (pp. 2-9). USA: IEEE. https://doi.org/10.1109/CESSER-IP.2019.00009.

Jeon, Y., & Seong, S. (2021). Robust recurrent network model for intermittent time-series forecasting. *International Journal of Forecasting*, *38*(4), 1415-1425. http://dx.doi.org/10.1016/j.ijforecast.2021.07.004.

Kormushev, P., Calinon, S., & Caldwell, D. G. (2013). Reinforcement learning in robotics: applications and real-world challenges. *Robotics*, *2*(3), 122-148. http://dx.doi.org/10.3390/robotics2030122.

Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. (2018). Sales-forecasting of Retail Stores using Machine Learning Techniques. In *Proceedings of the 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018*, 160–166. http://dx.doi.org/10.1109/CSITSS.2018.8768765

Ktenioudaki, A., O'Donnell, C. P., Emond, J. P., & do Nascimento Nunes, M. C. (2021). Blueberry supply chain: critical steps impacting fruit quality and application of a boosted regression tree model to predict weight loss. *Postharvest Biology and Technology*, *179*, 111590. http://dx.doi.org/10.1016/j.postharvbio.2021.111590.

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: deep learning really has pedagogical value. *Frontiers in Education*, *5*, 572367. http://dx.doi.org/10.3389/feduc.2020.572367.

Lorente-Leyva, L. L., Alemany, M. M. E., Peluffo-Ordóñez, D. H., & Araujo, R. A. (2021). Demand forecasting for textile products using statistical analysis and machine learning algorithms. In N. T. Nguyen, S. Chittayasothorn, D. Niyato & B. Trawiński (Eds.), *Intelligent Information and Database Systems. ACIIDS 2021. Lecture Notes in Computer Science* (vol. *12672*, pp. 181-194). Cham: Springer. http://dx.doi.org/10.1007/978-3-030-73280-6_15.

Lundberg, S. (2019). *GitHub - slundberg/shap: a game theoretic approach to explain the output of any machine learning model*. Retrieved January 24, 2022, from https://github.com/slundberg/shap

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles. *arXiv*, *1802.03888v3*, 1-9. https://doi.org/10.48550/arXiv.1802.03888.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv*, *1705.07874v2*, 1-10. https://doi.org/10.48550/arXiv.1705.07874.

Makridakis, S. (1988). Metaforecasting. *International Journal of Forecasting*, *4*(3), 467-491. http://dx.doi.org/10.1016/0169-2070(88)90112-4.

Marcilio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (pp. 340-347). USA: IEEE. https://doi.org/10.1109/SIBGRAPI51738.2020.00053.

Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2020). What makes an online review more helpful: an interpretation framework using XGBoost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(3), 466-490. http://dx.doi.org/10.3390/jtaer16030029.

Mitchell, T. M. (1997). *Does machine learning really work?* Retrieved January 24, 2022, from https://ojs.aaai.org/index.php/aimagazine/article/view/1303

Mitchell, T. M. (2006). *The discipline of machine learning*. Pittsburgh: Carnegie Mellon University.

Moore, J. D., & Swartout, W. R. (1988). *Explanation in expert systems: a survey*. University of Southern California Marina del Rey Information Sciences Inst.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: unbiased boosting with categorical features*. Retrieved March 2, 2022, from https://github.com/catboost/catboost

Raschka, S., & Mirjalili, V. (2015). *Python_machine_learning*. Birmingham, Reino Unido: Packt Publishing Ltd.

Ren, S., Chan, H. L., & Siqin, T. (2020). Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. *Annals of Operations Research*, *291*(1-2), 761-777. http://dx.doi.org/10.1007/s10479-019-03148-8.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv*, *1606.05386v1*, 91-95. https://doi.org/10.48550/arXiv.1606.05386.

Samek, W., & Müller, K.-R. (2019). Towards explainable artificial intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. Hansen & K. R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning. Lecture notes in computer science* (Vol. *11700*, pp. 5-22). Cham: Springer. http://dx.doi.org/10.1007/978-3-030-28954-6_1.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., & Young, M. (2014). Machine learning: the high-interest credit card of technical Debt. In *Proceedings of the SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)* (pp. 1-9). Google, Inc.

Seaman, B., & Bowman, J. (2021). Applicability of the M5 to Forecasting at Walmart. *International Journal of Forecasting*, *38*(4), 1468-1472. http://dx.doi.org/10.1016/j.ijforecast.2021.06.002.

Seeling, M. X., Scavarda, L. F., & Thomé, A. M. T. (2019). A sales and operations planning application in the Brazilian subsidiary of a multinational chemical company. *Brazilian Journal of Operations & Production Management*, *16*(3), 424-435. http://dx.doi.org/10.14488/BJOPM.2019.v16.n3.a6.

Shams Amiri, S., Mottahedi, S., Lee, E. R., & Hoque, S. (2021). Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption. *Computers, Environment and Urban Systems*, *88*, 101647. http://dx.doi.org/10.1016/j.compenvurbsys.2021.101647.

Sheshasaayee, A., & Logeshwari, L. (2018). Implementation of clustering technique based RFM analysis for customer behaviour in online transactions. In *Proceedings of the 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1166-1170). USA: IEEE. https://doi.org/10.1109/ICOEI.2018.8553873.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2016). *Inventory and production management in supply chains*. Boca Raton: CRC Press. https://doi.org/10.1201/9781315374406.

Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, *11*, 1-18.

Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275-294. http://dx.doi.org/10.1002/wics.1198.

Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, *46*(1), 411-419. http://dx.doi.org/10.1016/j.dss.2008.07.009.

Sutton, R., & Barto, A. G. (2018). *Reinforcement learning: an introduction* (2nd ed.). London: MIT Press.

Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine learning in predicting demand for fast-moving consumer goods: an exploratory research. *IFAC-PapersOnLine*, *52*(13), 737-742. http://dx.doi.org/10.1016/j.ifacol.2019.11.203.

Tirkolaee, E. B., Sadeghi, S., Mooseloo, F. M., Vandchali, H. R., & Aeini, S. (2021). Application of machine learning in supply chain management: a comprehensive overview of the main areas. *Mathematical Problems in Engineering*, *2021*, 1-14. http://dx.doi.org/10.1155/2021/1476043.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. Ch. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1-9. http://dx.doi.org/10.1016/j.simpat.2015.03.003.

Vogel, W., & Lasch, R. (2016). Complexity drivers in manufacturing companies: a literature review. *Logistics Research*, *9*(1), 25. http://dx.doi.org/10.1007/s12159-016-0152-9.

Wenzel, H., Smit, D., & Sardesai, S. (2019). A literature review on machine learning in supply chain management. In W. Kersten, T. Blecker & C. M. Ringle (Eds.), *Proceedings of the Hamburg International Conference of Logistics (HICL)* (pp. 413-441). Artificial Intelligence and Digital Transformation in Supply Chain Management. https://doi.org/10.15480/882.2478.

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., & Yu, Y. (2018). Taking human out of learning applications: a survey on automated machine learning. *arXiv*, *1810.13306v4*, 1-20. https://doi.org/10.48550/arXiv.1810.13306.

Zhang, B., & Ma, D. (2020). Flight delay prediciton at an airport using maching learning. In *Proceedings - 2020 5th International Conference on Electromechanical Control Technology and Transportation, ICECTT 2020* (pp. 557-560). USA: IEEE. http://dx.doi.org/10.1109/ICECTT50890.2020.00128.